Genes for Low-Molecular-Weight Heat Shock Proteins of Soybeans: Sequence Analysis of a Multigene Family

RONALD T. NAGAO,^{1*} EWA CZARNECKA,² WILLIAM B. GURLEY,² FRITZ SCHÖFFL,³ and JOE L. KEY¹

Botany Department, University of Georgia, Athens, Georgia 30602¹; Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida 32611²; and Fakultät für Biologie (Genetik), Universität Bielefeld, D-4800 Bielefeld 1, Federal Republic of Germany³

Received 22 July 1985/Accepted 19 September 1985

Soybeans, Glycine max, synthesize a family of low-molecular-weight heat shock (HS) proteins in response to HS. The DNA sequences of two genes encoding 17.5- and 17.6-kilodalton HS proteins were determined. Nuclease S1 mapping of the corresponding mRNA indicated multiple start termini at the 5' end and multiple stop termini at the 3' end. These two genes were compared with two other soybean HS genes of similar size. A comparison among the 5' flanking regions encompassing the presumptive HS promoter of the soybean HS-protein genes demonstrated this region to be extremely homologous. Analysis of the DNA sequences in the 5' flanking regions of the soybean genes with the corresponding regions of Drosophila melanogaster HS-protein genes revealed striking similarity between plants and animals in the presumptive promoter structure of thermoinducible genes. Sequences related to the Drosophila HS consensus regulatory element were found 57 to 62 base pairs 5' to the start of transcription in addition to secondary HS consensus elements located further upstream. Comparative analysis of the deduced amino acid sequences of four soybean HS proteins illustrated that these proteins were greater than 90% homologous. Comparison of the amino acid sequence for soybean HS proteins with other organisms showed much lower homology (less than 20%). Hydropathy profiles for Drosophila, Xenopus, Caenorhabditis elegans, and G. max HS proteins showed a similarity of major hydrophilic and hydrophobic regions, which suggests conservation of functional domains for these proteins among widely dispersed organisms.

All groups of organisms investigated undergo a response to high temperature referred to as heat shock (HS) (43). The HS response was first discovered in Drosophila melanogaster and has been studied in considerable detail in that organism (4, 43). This response is characterized by control mechanisms which are operative at the levels of both transcription and translation and is generally characterized by the induction of synthesis of a new set of proteins (HS proteins), decreased synthesis of most normal proteins, and the acquisition of thermotolerance to a nonpermissive (or lethal) HS temperature by prior exposure to permissive elevated temperatures. The induction of HS proteins is dependent on the transcriptional activation of a unique set of genes at the elevated or HS temperature. In D. melanogaster, four HS proteins in the range of 22 to 27 kilodaltons (kDa) and three high-molecular-weight groups of 68, 70, and 84 kDa are induced (4, 50). The 70-kDa class of HS proteins, arising from three genetic loci, represents a major proportion of total HS-protein synthesis in D. melanogaster and several other animal systems. In soybean, the high-molecular-weight HS proteins range from 68 to 110 kDa, and the small HS proteins are grouped between 15 and 27 kDa (22, 23, 26, 52).

The high-molecular-weight HS proteins seem to be highly conserved across a broad spectrum of organisms (21), whereas the small HS proteins show much more diversity (12, 13, 23, 46) in size and amino acid sequence. Sequence conservation between plants and animals among the genes encoding the high-molecular-weight HS proteins is evident by the cross-hybridization of plant genomic clones and poly(A)⁺ RNAs with *Drosophila* HS cDNA clones (47; J. Roberts, unpublished data) and by cross-reactivity of anti-

Because of the distinct abundance and complexity of low-molecular-weight HS proteins in plants, our efforts have concentrated on the analysis of mRNA induction and the isolation and characterization of genes for this class of HS proteins. With HS-specific cDNA clones as hybridization probes, HS mRNAs corresponding to the small HS proteins are detectable within 3 to 5 min after HS (44). Liquid hybridization studies indicate that about 20 of these mRNAs accumulate to 20,000 copies each per cell within 2 h at 40°C (44). Considerable homology among this class of HS mRNAs is demonstrated by two-dimensional hybrid select and translation analyses with HS cDNA clones. Clone pCE53 or pFS2005 yields 13 proteins of 15 to 18 kDa, whereas pFS2019 yields a single 18-kDa protein within the same group (24, 44). Similar analyses with HS cDNA clone pFS2033 show three proteins of 21 to 24 kDa (23, 44), whereas pCE54 hybrid select translates into five 27-kDa proteins (12). These analyses and Southern hybridization analyses (23, 45) indicate the existence of several HS gene families within the low-molecular-weight group of proteins in soybean, with some families comprising only a few members and others up to 13 closely related proteins.

bodies (21). The high-molecular-weight HS proteins of plants, in contrast to those of *D. melanogaster* (14, 15, 29), represent a relatively small fraction of total HS-protein accumulation (e.g., soybean [22, 23, 25, 26]). The major HS-protein accumulation in plants is, instead, represented by a complex group of about 20 15- to 18-kDa proteins and approximately 10 20- to 27-kDa proteins based on radioactive amino acid incorporation and Coomassie-stained gel analyses (25). Although all plant species investigated to date synthesize a complex array of 15- to 27-kDa HS proteins, the electrophoretic patterns of these proteins vary among species (22).

^{*} Corresponding author.

As in other systems (4, 43), the HS genes of soybean are also induced to various degrees by such agents as arsenite, cadmium, and various environmental stresses (12). We see no evidence for the expression of soybean HS proteins in the 15- to 24-kDa groups under normal developmental or hormone-induced states. However, the 27-kDa group of HS proteins in soybean is expressed at control temperatures (28°C) and is induced 5- to 20-fold by elevated temperatures and a variety of other stresses (12). A similar type of enhanced expression has been observed for some of the *Drosophila* genes (36).

In this report and one by Czarnecka et al. (13), we present the DNA sequence analyses and transcript mapping results for three closely related genes within the 15- to 18-kDa group of HS proteins. A comparison of DNA sequences in the 5' flanking region of the soybean genes with the corresponding regions of *Drosophila* HS genes reveals a striking similarity between plants and animals in the presumptive promoter structure of thermoinducible genes. The relationship of the low-molecular-weight HS proteins of soybean to the four small HS proteins of *D. melanogaster* is characterized by a comparison of the deduced amino acid sequences and by analysis of hydropathy plots.

MATERIALS AND METHODS

Restriction endonucleases, T4 DNA ligase, and DNA polymerase I large fragment were obtained from Bethesda Research Laboratories, Inc. (Gaithersburg, Md.), New England BioLabs, Inc. (Beverly, Mass.), and New England Nuclear Corp. (Boston, Mass.). Calf intestinal phosphatase was purchased from Boehringer Mannheim Biochemicals, (Indianapolis, Ind.). Polynucleotide kinase was obtained from Pharmacia, Inc. (Piscataway, N.J.). $\left[\alpha^{-32}P\right]dNTP$ and $[\gamma^{-32}P]ATP$ were purchased from New England Nuclear. Chemicals used for DNA sequencing were from vendors recommended by Maxam and Gilbert (31). X-ray film, X-Omat AR-5, was supplied as long rolls by the Eastman Kodak Co. (Rochester, N.Y.). Acrylamide was purchased from Kodak and purified as described by Maniatis et al. (30) with the inclusion of a charcoal decolorization step (5 g/liter). All other chemicals were reagent grade unless otherwise stated.

Soybean genomic DNA library. Total soybean (*Glycine* max var. Corsoy) DNA was partially digested with *MboI* and ligated into the *Bam*HI site of the cloning vector λ_{1059} (20). The soybean λ_{1059} genomic library was constructed by J. Slightom and Y. Ma, Agrigenetics Advanced Research Laboratory, Madison, Wis. (10, 48). Screening was as described by Nagao et al. (34) with purified HS cDNA inserts labeled by nick translation (30) with ³²P. Lambda clones characterized in this study were designated Gmhs λ L (λ L), Gmhs λ M (λ M), and Gmhs λ E (λ E). HS genes contained within these clones were subcloned into pUC9 (51) and designated *Gmhsp17.6-L* (*G. max* HS protein, 17.6 kDa), *Gmhsp17.5-M*, and *Gmhsp17.5-E* (corresponding plasmid clones are abbreviated pL, pM, and pE, respectively).

Subcloning. Southern hybridization with cDNA insert (pFS2005 [44]) was used to identify appropriate fragments from an *Eco*RI restriction digestion for subcloning (30) into pUC9 (host strain JM83) (51). An approximately 2.4-kilobase (kb) *Eco*RI fragment isolated from Gmhs λ M was subcloned and designated pM/EE2.4 (abbreviated pM). A 1.7-kilobase *Eco*RI insert subcloned from Gmhs λ L was designated pL/EE1.7 (abbreviated pL).

DNA sequence determination. The reactions for DNA

sequence determination were those described by Maxam and Gilbert (31), except that formic acid was used for the A+G reaction. Protocols for sequencing procedures and the use of long (100-cm) sequencing gels were as described by Barker et al. (5). Computer analyses of DNA and protein sequences were performed with computer programs made available by J. Pustell and F. Kafatos (41) and modified for an HP 1000 (Hewlett-Packard Co., Palo Alto, Calif.) by M. Clegg and J. McClendon (University of Georgia, Athens).

Nuclease S1 hybrid protection mapping. A BamHI site located within the coding region of all three HS genes served as a reference point for mapping the 5' and 3' termini of the HS transcripts according to the procedure of Favaloro et al. (17). The 5' terminus of each gene was mapped with an appropriate DNA restriction fragment as hybridization probe that extended several hundred base pairs (bp) upstream from the 5' end-labeled (31) BamHI site and included the presumed cap site and TATA-like regions. The 3' termini were mapped with DNA restriction fragments that extended downstream from the 3' end-labeled (30) BamHI site to include a few hundred bp distal to the deduced translational termination codon. Poly(A)⁺ RNA (1 μ g) isolated from control (28°C) and HS (40°C) soybean seedlings according to Czarnecka et al. (12) was hybridized (for 12 to 18 h) to end-labeled probe DNA at various temperatures ranging from 42 to 56°C to determine the optimum for hybrid formation. Nuclease S1 digestions were performed at 50 to 200 μ /ml for 30 min at 15°C. Protected hybrids were precipitated with isopropanol and detected by autoradiography after fractionation by electrophoresis on 6% polyacrylamideurea sequencing gels.

RESULTS

Characterization of HS genomic clones. Soybean HS genomic clones were isolated from a λ_{1059} library by screening with HS cDNA inserts as described above. Approximately 40 independent clones were initially purified, each containing from 15 to 20 kilobases of genomic sequences. Our general strategy for mapping HS genes within large genomic inserts was first to identify the region of cDNA homology and start DNA sequencing in both directions from that point. Once restriction sites internal to HS mRNA were identified by their presence in the cDNA-homologous portion of the gene, nuclease S1 hybrid protection mapping was performed to position the 5' and 3' termini of the transcript. In several cases, the cDNA used originally to select a particular lambda clone did not show 100% homology to the genomic DNA. This apparent discrepancy is explained by cross-hybridization between the various members of closely related multigene families as was demonstrated previously in cDNA hybrid selection and in vitro translation studies (12, 23, 44). In all cases discussed here, the cDNA was colinear, with allowances for mismatch, to genomic sequences, indicating the absence of introns in these HS genes. The corresponding proteins for these genomic clones, based on cDNA hybrid select and translation, ranged in molecular weight from 15 to 27 kDa. The sequences of two selected representatives of the 15- to 18-kDa gene family are presented here.

Figure 1 shows a partial restriction map and strategies used for the determination of DNA sequences of genomic subclones pL and pM. The nucleotide sequence of each gene, along with its predicted amino acid sequence, is presented in Fig. 2. For comparative purposes, pE, the sequence of another HS gene in this size class, reported recently by Czarnecka et al. (13), will also be discussed. For



FIG. 1. Restriction map and sequencing strategy for *Eco*RI subclones pL/EE1.7 and pM/EE2.4 containing the HS genes *Gmhsp17.6-L* and *Gmhsp17.5-M*, respectively. Double line of restriction map indicates soybean genomic DNA; single line indicates plasmid vector pUC9. The strategies of restriction sites used for nucleotide sequence determination are shown below. The locations and orientations of protein coding regions are indicated above the restriction maps.

each of these genes there is close agreement between the molecular weight of the protein predicted from the deduced amino acid sequence and the molecular weights observed in hybrid selection and in vitro translation experiments (25, 44).

5' Terminus of HS mRNAs. The 5' termini of the HS mRNAs were positioned on the genomic sequences by nuclease S1 hybrid protection mapping with end-labeled DNA hybridization probes (17). Multiple 5' termini were observed for both pL and pM transcripts (Fig. 3), whereas the closely related gene pE has a single 5' terminus (13). For pL, there were two protected fragments of approximately equivalent intensity with lengths of 169 and 172 bp. S1 protection analysis for pM also yielded two fragments with lengths of 165 and 170 bp. The initiation sites for each of these genes are composed of short direct repeat sequences, both of which are used for initiation, except in the case of pE, in which only the TATA distal site is utilized. The

redundant initiation sites for pL, pM, and pE are CATCATC, AAACGAAACG, and TCGTCCTCGTC, respectively, with two adenines occupying position \pm 1 for pL and pM transcripts and a single guanine for the pE transcript. The 5' leader sequences for these three gene transcripts are predicted to range from 82 to 96 bp. Some comparative features of the low-molecular-weight HS genes and proteins are presented in Table 1.

Analysis of the 3' nontranslated region. DNA sequences of the 3' nontranslated portion of the genes showed regions of extensive homology among the soybean genes based on homology matrix analyses (data not shown). The most notable regions of common homology reside around nucleotides 559 to 631, 651 to 692, and 721 to 730. Comparison of the 3' end of soybean HS genes showed no consistent homology with the corresponding region of *Drosophila* HS genes.

						-320	-310
-300	-290	-280 GTITITGTTT	-270 AAGTTACTGT	-260 ACTGTGGGCC	-250	-240 TAGATCAAAG	-230 TAGTAATAAT
-220 AATATTGATT	-210 AAATGATATA	-200 TATATATATA	-190 TATATATATA	-180 TATAT <u>CTAGA</u>	-170	-160 AGACTAGCTA	-150 GAACGTACGT

ACGT AATAT ATTOGTOTO AGAAGTOCTO AAGTITATOO AATCATOTAA AACTOCTAAA ATAGCAAACA ACATTATATT GTAAACAATA -50 -70 -70 -10 10 <u>TITITCIGGA</u> ÁCATACAAGA GTATCCTITC ACTICC<u>TITA AATA</u>CCTCGA GTGTCCCCCAT TGACATCATC AAACAAGAGA 20 30 40 50 60 70 80 AGAGTTACAG AATTTCCTGT TTACGATCTC ATTACAATTT TGCAACTTTC AAAGCTTATT AGCTAAAGTA ACATCAAAAG

100 110 120 130 140 150 ATG TCA TTG ATT CCA AGT ATT TTC GGT GGC CCA AGG AGC AAC GTG TTC GAT CCA TTC TCA Met Ser Leu IIe Pro Ser IIe Phe GIy Giy Pro Arg Ser Asn Vai Phe Asp Pro Phe Ser 180 170 180 190 200 210 CTC GAT ATG TGG GAT CCC TTC AAG GAT TTT CAT GTT CCC ACT TCT TCT GTT TCT GCT GAA Leu Asp Met Trp Asp Pro Phe Lys Asp Phe His Vol Pro Thr Ser Ser Vol Ser Alo Giu 220 230 240 250 270 AAT TCT CCA TTT GTG AAC ACA CGT GTG GAT TGG AAG GAG ACC CAA GAG GCA CAC GTG CTC Asn Ser Alg Phe Vol Asn Thr Arg Vol Asp Trp Lys Glu Thr Gin Glu Alg His Vol Leu 280 280 300 310 320 330 AAG GCT GAT ATT CCA GGG CTG AAG AAA GAG GAA GTG AAG GTT CAG ATT GAA GAT AGG Lys Alg Asp lie Pro Gly Leu Lys Lys Glu Glu Vol Lys Vol Gin lie Glu Asp Asp Arg 340 350 340 370 340 360 GTT CTT CAG ATT AGC GGA GAG AGG AAC GTT GAG AAG GAA GAC AAG AAC GAC ACG TGG CAT Voil Leu Gin IIe Ser Giy Giu Arg Asn Voil Giu Lys Giu Asp Lys Asn Asp Thr Trp His 400 410 420 430 440 450 CGC GTG GAC CGT AGC AGT GGA AAG TTC ATG AGA AGG TTC AGA TTG CCA GAG AAT GCA AAA Arg Val Asp Arg Ser Ser Gly Lys Phe Met Arg Arg Phe Arg Leu Pro Glu Asn Ala Lys 440 570 480 490 500 510 GG GAG CAA GTA AAG GCT TGT ATG GAA AAT GGG GTT CTC ACT GTT ACT ATT CCA AAG GAA Vel Giu Gin Vel Lys Alg Cys Met Giu <u>Ann Giy Vel Leu Thr</u> Vel Thr ile Pre Lys Giu 520 530 540 550 560 570 GAG GTT AAG AAG TCT GAT GTT AAG CCT ATA GAA ATC TCT GGT TAA ACTTGGTTTC ACTGAAAATC GIU Val Lys Lys Ser Asp Val Lys Pro IIe Giu IIe Ser Giy ---GTGAGAGCTT TTAAATTTGC TTTGTTGTAA TAAGTGTCCT TTGTCTTGTG TTCCAATGGT GATTTTGAGA AAGATCATAC eso ero eso eso 700 710 720 730 AATTGTGCCT TGTGTTGTTG TGCAAGTGTA ATTGAAGTGA ATAAAAAATT AACACCTGCT TTCAGAAAAT TTTGCTGTGT 740 750 780 770 780 780 800 810 GTCATTGTCA TCGAATATGT GATGTAGGCA AGAAATAGAC CGTGAAAATA ATATCTGACA TITGGCTAAT TGCTTTTGTT ATGCTGAGAC ACTCTATGTG AAATAACTGC ATTTATCATG TTCCATCTTC TTAATACAAG AAGTCAATAC CAATGTCTTA CCAMATTANG ATAACAGGTT GATTTGGACT CATCANAGTG CAGCCCTTTA TTTGGACTCA TCAAAGTGCA GCACTANAGG GTTTTGTTAA CTAGCAAGTT CAGAGCATCA TTTAAGTAAT TAAAAGAAAA AATATTAAAT ATATAAATCA TAAGATGATA 1060 1070 1080 1090 1100 1110 1120 1300 TCAAAAAATT CATGAACAGT CTCTTCATTT TTTTTCAATA AAAATATTTT TATTTTAATT TTTTAAAATA ATATCCTCAT 1170 AACATTEGETT TAACTECECAA GTTTAAAATT TACTAGTECT AGATAAATTE TETAAGATAA TETATAGATA AAAATAAGAT AMATTAGAMA ATTITTAAGG AGAGATTITT TITTATAAMA ATTAGGTATA TOTATTGGTT TAGTTTACA GAGAMATATA 1300 1310 1320 1340 1350 1350 1350 1350 1350 1350 1350 1380 GAATTCT

MOL. CELL. BIOL.

000

- 760 ATAATGTTTA TGAGGCTATT GCTATATCCA TTCCTCTTGT TGCTAAGTAC ACCCTTACCT TTTTTCTAAA CAAATACTTA AATAACCTIT TCCCTTCAAA CCTTCTCTCT TGCGCACTCC CTTTTCCCCCC ACAACAACAA ACTTCTCTTT TTCATAGCAA -600 -620 CAACACATTT TITICICITIT CATCACATTC TCACCICICI TITCATICCC TICAATCICA TGCICIGGTC TCCCTTCTCC -55C -54C -53C -52C -51C -50C -49C -49C CCTGTTTAG TCTTCCTTGCTT TAGTGTTAGG ATTTTTTCAT <u>CTTCAAACTT CAAGT</u>TGTTG TAGTGGACCA AGTCAAATTG -450 -440 - 430 -420 CTGAGATTGA AGGACTAATA CAAAATATAT GTTTTTTCCA CTTCAGTTCC TCCACCATAG TCTTAGTATT CTGCTTGAAT - 280 ACACATTITT TITAATATTC IGAAAAAATAT TITICAGAAC ACAACAATAT TICAGAATTT ATAGGTACAA AGATTITAAT -220 -210 -200 ANAAAAGGAT GGTGAATATA GCAAAAAGCCT ATTTATGAAC GATATCAACC AGAACTAGAA CAAGAAAAAAT AAATGCACTA GAACCITCGT ACACGGAGTG GAGAAGTCCA GAAGTTITIA TAGAATCATT TGAAACTGGT AAAACCAACC AAATTGCAAA 10 20 30 40 50 60 75 80 AAAAGAGTTA CAAAAGTAACC TGTATACGAT CICATTTTGA TCTCCCAAGT TTCAAATCIC GCGAATAAAT ATATCAAAAAG po 100 110 120 130 140 ATG TCT CTG ATT CCA AGT ATT TTC GGT GGC AGA AGG AGC AAC GTG TTC GAT CCT TTC TCC Met Ser Leu IIe Pro Ser IIe Phe GIy GIy Arg Arg Ser Asn Vol Phe Asp Pro Phe Ser CTC GAC GTG TGG GAT CCC TTC AAG GAT TTT CAT TTT CCC ACT TCT CTT TCT GCT GAA AAT Leu Asp Val Trp Asp Pro Phe Lys Asp Phe His Phe Pro Thr Ser Leu Ser Ald Glu Ash 230 240 270 280 290 30C 310 32C GCT GAT ATT CCA GGG CTG AAG GAG GAG GTG AAG GTG CAG ATT GAA GAT GAT AGG GTT Ale Asp lie Pro Giy Leu Lys Lys Giu Giu Vol Lys Vol Gin lie Giu Asp Asp Arg Vol 330 340 350 340 370 380 CTT CAG ATA AGC GGA GAG AGG AAC CTT GAG AAG GAA GAC AAG AAC GAC AGG TGG CAT CGC Leu Gin He Ser Giy Giu Arg Ash Leu Giu Lys Giu Asp Lys Ash Asp Thr Trp His Arg 470 GÃO CAA GTG AAG GCT TCT ATG GÃA AAT GGT GTT CTC ACA GTT ACC GTT CCC ÀAA GAA GAG GIU GIN VOI Lys AIO Ser Met GIU <u>Asn Giy Voi Leu Thr</u> Voi Thr Voi Pro Lys Giu Giu 680 ATTGTGAMAT GTANCTGCTT ANGTGGTANT TGTGTGCTTG TGTANTTCAT GTGANTANAG TCCAGCCTTC AGAAAATACT ATAATTTCCT ATATCACAAG TGTCTGATAT GTGATGTAGA CAAAAAATA GATTGTGTAA TGTGAATTGT AATTGCCCTA TAGCGTGCAT TTGTTATAGC CGAGAGACTT TCTCTCTATG TTCATAAACT CCACCTCCCA GTAAGTGTAT AATATTCCAT CTTAATAATG AAAATAAATA GACGATTTAG TGTGTTCCCA TGCATAACTA AAGACTATAG AGGATTTGAT ATTCTCTTCC 900 990 1000 1010 1020 103C 1040 1050 AAGGGCAAGA GTAAATIGAA AATAAGACTA GAACCAACTT TITIGTCTIG ATTITTAATA AACAAATGTA CAACCGAGGA 1070 1100 1110 TATATAGACT GTGGTCATCA ACAGCATAAA ATCACTTAGA AGTTGGAATA GCTAGCTTGC ATTATTTAGG TTGTAAAGAC 1140 1150 1160 1170 1180 1190 1200 1270 CCTGCAATAA CCTTGAAAAT TTTGAGGCTT TGGATACTCA TCTGTTTGAT TAATTATCTA AATTITTAAGA CTTATATTTA 1220 1230 1240 1250 :260 1270 TTTATACGCA AGTAGGTGTG GAAATGAGTC AAGTTAAGTC AAGCTTTATT AGGCTTGAGC TCGGCTTGAG TTGAATACGT 1330 :340 AAGGCTTGGG TTTGACTTAA TTAACCTGAT CTAATTAGAC TTTTTTAAAG GATCTCTAGA TCTCGATCTT ACATAAAAGT :410 1420 :430 1390 1400 1440

CTGGCTTGAC CCACGAGCCT ATTTAAAAGC TTGCTTAAAG ACGTCTTTGA TTAATTAATT ATTTTAAAAC CTAGTGAAAT 1460 1470 1480 1500 1510 520 530 ACTAACTAAA AAAAGAAACT TATAAATTIC ATGTAAATAA TGTACAAATC CAAAAATAAT TGATAAACAA AATCATATTG 1560 1570 1590 AATTCACTGG CCGTCGTTTT ACAACGTCGT GACTGGGAAA ACCCTGGCGT TACCCAACTT AATCGCCTTG CAGCAC

FIG. 2. The complete nucleotide sequences and deduced amino acid sequences of Gmhsp17.6-L (A) and Gmhsp17.5-M (B). For consistency, nucleotides are numbered from the distal start site for transcription, but both cap sites are indicated by arrows. The Drosophila

HS consensus is underlined with asterisks denoting nucleotides homologous to the core inverted repeat at 90%. The TATA-like motif is bold underlined. Arrows denote termini of mRNA as determined by nuclease S1 hybrid protection analysis. The highly conserved amino acids found in the hydrophobic region of Drosophila and soybean low-molecular-weight HS proteins is underscored.



FIG. 3. Nuclease S1 mapping of 5' termini of soybean HS genes Gmhsp17.5-L (A) and Gmhsp17.5-M (B). An internal BamHI site was 5' end labeled and used as a hybridization probe with 1 µg of poly(A)⁺ RNA isolated from seedlings incubated at control (28°C) and HS (40°C) temperatures. Lanes: 2, 4, and 6, 28°C RNA; 3, 5, and 7, 40°C RNA. Temperatures of hybridization were: 42°C for lanes 2 and 3, 45°C for lanes 4 and 5, and 48°C for lanes 6 and 7. Bands (denoted by arrows) indicating 5' termini are at 172 and 169 bases for Gmhsp17.5-L (A), and 170 and 165 bases for Gmhsp17.5-M (B). The high-molecular-weight (500-bp) band in lane 5 represents intact probe. Sizes (in bp) of $\phi X174$ -HaeIII DNA marker fragments are indicated in the margins.

The location of 3' termini for transcripts of pL and pM were established with nuclease S1 hybrid protection studies. Multiple 3' termini were observed for RNA homologous to both genes. Figure 4 demonstrates the utilization of two major (575 and 550 bases), two moderate (590 and 560 bases), and three minor (740, 710, and 660 bases) 3' termini corresponding to transcripts from Gmhsp17.6-L. Mapping of the 3' termini of transcripts homologous to Gmhsp17.5-M resulted in major bands at 580 and 570 bases, a moderately abundant band at 706 bases, and five to six minor bands from 780 to 520 bases. The presence of numerous minor bands is likely due, at least in part, to the complexity and interrelatedness of this multigene family, where some homology has been demonstrated in the 3' nontranslated region. The possibility of hybridization bands arising from transcripts from different alleles of the same gene cannot be eliminated, but to reduce confusion arising from various degrees of cross-hybridization, we altered the criterion of RNA-DNA duplex formation by conducting hybridizations for nuclease S1 mapping at various temperatures. Autoradiographic bands that did not disappear or change in relative intensity with increasing stringency of hybridization were used to establish the location of presumptive 3' termini or poly(A) addition sites. Thus, by this criterion, bands of 410, 395, and 380 for pL and 430, 420, 410, and 400 for pM are very likely the result of hybridizations with related but different members of this complex family of mRNAs. Many of the 3' termini of pL and pM are located within a distance of 35 nucleotides downstream from a sequence similar to the mammalian consensus polyadenylation signal AATAAA (40) (for example, pL bands at 740, 710, and 660 bases; pM bands at 706 and 650 bases); however, many termini have no apparent correlation with the presence of a consensus-like polyadenylation sequence.

The 5' flanking sequences. Once the 5' terminus of the RNA was determined, the 5' flanking sequences were

searched for putative eucaryotic transcriptional regulatory elements. For the HS genes of D. melanogaster, the bestcharacterized promoter elements include an AT-rich motif known as the TATA box (Goldberg-Hogness box) and the HS consensus sequence CTgGAAtnTTCtAGA (38). A comparison of the 5' flanking sequences of four sovbean HS genes (including HS6871 [46]) revealed a striking degree of similarity in sequences between -23 and -72 bp, which include both a TATA-like motif and an upstream region with considerable homology to the Drosophila HS consensus. The TATA-like motif, TTTAAATA, was present in each of these soybean HS genes from 27 to 31 bp upstream from the transcription start sites and from 110 to 132 bp 5' to the initiation codon for translation. Upstream, 31 and 41 bp from the 5' end of the TTTAAATA motif, are the 5' ends of two overlapping 15-bp sequences with homology to the Drosophila HS consensus promoter (38). Homology is 79% for both the overlapping consensus sequences of pE; 64 and 79% for the upstream and TATA-proximal overlapping consensus sequences, respectively, in pL; 71 and 64% respectively in pM; and 64 and 71%, respectively, in HS6871 (46). The TATA-proximal HS consensus also demonstrated high homology to the HS core inverted-repeat consensus sequence identified in D. melanogaster (CTnGAAnnT TCnAG), with 90% homology for pE and pL and 80% for pM and HS6871. Secondary regions of high homology to the Drosophila HS consensus also occurred much further upstream from the TATA-proximal, or primary, HS element in each of the soybean genes. The HS6871, pL, and pM genes all had secondary HS elements with 90% homology to the core inverted repeat. These secondary consensus sequences were centered 86, 95, 106, and 137 bp upstream from the 5' end of the TATA-like motif in HS6871; 145 bp upstream in Gmhsp17.6-L; and 472 bp from TATA in Gmhsp17.5-M. Gene Gmhsp17.5-E contained a single secondary consensus sequence 332 bp upstream from TATA with 80% homology to the Drosophila core inverted repeat.

Strong similarities among the soybean low-molecularweight HS protein genes became more evident when approximately 300 nucleotides 5' to the translation start codon of the soybean HS-protein genes were aligned (Fig. 5). The overall sequence homologies were: 68% between pM and pL, 74% between pE and pM, and 74% between pE and pL. A region of even higher homology (84 to 88%) lay between the 42 nucleotides from the 5' end of the TTTAAATA motif through the 5' end of the HS consensus sequence. The region of highest sequence heterogeneity was located between the 3' end of the TTTAAATA motif and the CAP site. Over this 24-nucleotide sequence, the homologies between pE and pM, pL and pM, and pL and pE were only 29, 42, and 67%, respectively.

Alternating purine-pyrimidine stretches have the potential under certain conditions to form Z-DNA (35). Several regions of alternating purines and pyrimidines were present in the 5' flanking sequences of genes Gmhsp17.6-L and Gmhsp17.5-E. In Gmhsp17.6-L, a group of 15 alternating pairs of AT occurred adjacent to a secondary HS consensus centered at position -198. In Gmhsp17.5-E, short clusters were centered at positions -130 and -106 immediately upstream from a sequence showing 78% (11 of 14) homology to the simian virus 40 enhancer core.

Analysis of deduced amino acid sequences. The soybean gene sequences presented here each contained an uninterrupted open reading frame starting at the first ATG 3' to the TTTAAATA motif. The molecular weights deduced from the single, uninterrupted open reading frames of these genes

		Sequences:		% Base		Protein-co	ding sequence:	
Organism and gene or protein	TATA box- initiator codon distance (bp)	Estimated leader sequence length (bp)	% A+T content of first 200- bp extragenic sequence	composition of leader sequences (A+T content)	Length of open reading frame (bp)	Length of encoded polypeptide (residues)	Mol wt of unmodified polypeptide chain	Termination codon
D. melanogaster ^a								1.0
hsp22	284	ND^{b}	57	70	525	174	19,705	TAG
hsp23	146	119 ± 3	47	67	561	186	20.630	TAG
hsp26	215	178 ± 3	66	67	627	208	22,997	TAA
hsp27	151	ND	61	69	642	213	23,620	TAA
Sovbean ^c								
Gmhsp17.5-M	110	88 + 93	68 ^d	68	459	153	17,544	ТАА
Gmhsp17.6-L	123	93 + 96	69 ^e	70	462	154	17,570	ТАА
Gmhsp17.5-E	113	82	62	65	462	154	17.533	TGA
HS6871	132	104	64	65	459	153	17.345	TAA

TABLE 1. Some features of small HS genes and proteins

were 17.6, 17.5, 17.5, and 17.3 kDa for genes *Gmhsp17.6-L*, *Gmhsp17.5-M*, *Gmhsp17.5-E*, and *HS6871*, respectively. The deduced molecular weights of these four genes are consistent with experimental determinations of molecular weights of 15 to 18 kDa based on in vivo-labeled HS proteins, in vitro translation of HS poly(A) RNA, and HS cDNA hybrid select and translated proteins (26, 44). This agreement between the deduced and observed molecular weights of these proteins, the comparison of partial-length cDNA sequences with genomic DNA sequences, and the



FIG. 4. Nuclease S1 mapping of the 3' termini of soybean HS genes *Gmhsp17.5-L* (A) and *Gmhsp17.5-M* (B). An internal *Bam*HI site was 3' end labeled and used as a hybridization probe with 1 μ g of poly(A)⁺ RNA isolated from seedlings incubated at control (28°C) and HS (40°C) temperatures. RNA samples are designated above the lanes as control (C), heat shock (HS), and yeast tRNA (T). Temperatures (in °C) of hybridization are indicated below the lanes. Major 3' termini are indicated by bands (denoted by arrows) at 575 and 550 bases for *Gmhsp17.5-L* (A) and at 580 and 570 bases for *Gmhsp17.5-M* (B). Sizes (in bp) of ϕ X174-*Hae*III DNA marker (M) fragments are indicated in the margins.

transcript mapping studies indicate that these four HS genes do not contain introns.

Amino acid sequence comparisons of the four soybean HS proteins revealed homologies of greater than 90% (Fig. 6). Of the nucleotide changes observed in the open reading frame, approximately two-thirds were silent substitutions. For example, of the 42 nucleotide differences between Gmhsp17.5-M and Gmhsp17.6-L, 28 were synonymous substitutions with 14 nucleotide changes leading to amino acid changes. Thirty-nine nucleotide changes were observed when genes Gmhsp17.5-M and Gmhsp17.5-E were compared; of these, 26 are synonymous substitutions, and 13 lead to amino acid changes. A comparison of Gmhsp17.5-E with Gmhsp17.6-L showed 31 nucleotide changes, of which 16 are synonymous substitutions, and 15 cause amino acid changes. Comparison of Gmhsp17.5-E with HS6871 showed 34 synonymous substitutions out of 47 total changes. In addition, a single amino acid deletion is predicted in the sequences of Gmhsp17.5-M and HS6871. Tyrosine is not encoded in the proteins of Gmhsp17.5-E and HS6861, whereas cysteine is only encoded in the Gmhsp17.6-L sequence.

A comparison of the deduced amino acid sequences of six proteins in the small HS family from C. elegans (42), D. melanogaster (18), and Xenopus laevis (7) with the four low-molecular-weight HS proteins of soybean is presented in Fig. 7. Only the carboxy-terminal half of the proteins is shown, since a significant level of homology was not seen in the amino-terminal regions. Spaces represent adjustments in alignment to maximize homology. There were five positions where the same amino acid is used in all 10 of the proteins examined. Of 92 positions, 15 (16%) in the carboxy-terminal portion of the soybean genes had the identical amino acid found in at least three other proteins. Of 92 soybean positions, 38 (41%) were identical, with at least one protein from the other organisms. Separate comparisons of the composite amino acid sequences of C. elegans and X. laevis over the region presented with the composite of the four proteins of D. melanogaster show 39 and 37% homology, respectively. A similar comparison of the composite soybean amino acid sequences with D. melanogaster reveals a somewhat lower homology of 28%. Maximum alignment was obtained by assuming that a deletion occurred near the carboxy terminus in the soybean proteins. These results demonstrate significant homology in primary amino acid sequence between the

				Bas	e composit	ion (%)		E	Base compo te	sition (in 1 rmination (00 bp) 3' to the odon				
Acidic amino acids (%)	Basic amino acids(% Arg + Lys)	Basic amino acids (% Arg + Lys + His)	A	G	С	Т	G+C	A	G	C	Т	A+T			
16.7	12.1	14.4	22	32	28	17	58	41	13	13	33	74			
14.0	11.3	14.0	22	31	27	20	60	32	15	19	34	66			
12.5	12.5	16.8	22	29	30	19	59	30	9	21	38	68			
13.1	12.7	18.8	22	31	29	18	60	43	22	17	16	59			
18.3 17 4	15.1	17.1 17.4	27 30	29 28	19 17	25 25	48 45	34 36	17	14 12	34 38	68 74			
17.4	16.1	18.0	27	28	19	26	47	36	16	15	33	69			
17.0	16.4	17.8	28	29	18	25	47	31	18	13	38	69			

TABLE 1-Continued

^a Drosophila data from Southgate et al. (49).

^b ND, Not determined.

'Soybean data for Gmhsp17.5-E and HS6871 from Czarnecka et al. (13) and Schöffl et al. (46), respectively. ^d M clone 600 bp of 5' = 70% A+T; M clone 957 bp of 5' = 69% A+T.

" L clone 417 bp of 5' = 70% A+T.

low-molecular-weight family of HS proteins in soybean and the class of proteins present in animals known as the small HS proteins (18).

Hydropathy plots provide a means to evaluate common structural features of distantly related proteins (27). Hydropathy profiles of the four soybean proteins were nearly identical over the entire length because of the high

degree of amino acid homology found among members of this closely related family of HS proteins. A comparison of a representative (Gmhsp17.6-L) of the soybean HS proteins with small HS proteins from other organisms gave a more complete view of the similarity in structural features among members of this diverse class of proteins. The hydropathy profiles shown in Fig. 8 present representative proteins from

	-210	-200	-190	-180	-170	-160	-150	-140	
Gmhs 17.5M	TATAGCÁAAAG	CCTATTTATGA	ACGATATCAA	CCAGĂACTA	GAACAÁGAAA	AATAAÁTGCAC	TAGAÁCCTT	CGTACÁCG	
Gmhs 17.6L	TTAAATGATATA	TATATATAT AT	ΓΑΤΑΤΑΤ ΑΤ	ATATATCTA	GAAGGTTGTA	₿ <u>GA</u> -G-CTAG-	G-A	TT	
Gmhs 17.5E	A-AGAATTTCTA	TA-GA-G	-AAGA	AACCT	ATGT-TTT-T	G-ATG T-	C G-G	GAAGA-AA	
Gmhs 17.5M	-130 GAGTGGAGAAGT	-120 CCAGAAGTTTT	-110 TATAGAATCA	-100 TTTGAAACT	-90 GGTAAAA	a- AAADDAADD	D TTGC AA	-70 A <u>CACGATI</u>	
Gmhs 17.6L	Ţ	A-	· C	-C-A	-CTAG	CAAAC-ATT	A-AGT	-CAATAII	
Gmhs 17.5E	ATAAAT-TTG	ATGTGTAG-AAA	CACCT	-GC-T-CA-	GTGGAGA	ATT	AAA	G <u>taggatt</u>	
Gmhs 17.5M	- 0 <u>111<u>ctgga</u>acgt/</u>	-50 ACACGATTATCO	-40 CTTTCACTTA	-30 CIIIAAAIA	-20 CATCACGATT	-10 AGTCAGAA <u>A</u> AA		10 GAAGAAAA	
Gmhs 17.6L	ITICTGGAACAT	<u></u>		- <u>TTTAAATA</u>	-CGA-TG-	CCCTTG-CA	тсатс	G	
Gmhs 17.5E	<u>tttctcca</u> acat	<u></u>	c	-IIIAAAIA	-CGTA-	ссс-ттсстсс	тс <u>с</u> тс		
Gmhs 17.5 M	20 GAGTTACAAAGT	30 ACCTGTATACO	40 SATCTCATTT	50 TGATCTCCC	eo AAGTTTCAAA	ፙ TCTC GCGAAT	ao AAA T AT	ATCAAAAG AT	۱G
Gmhs 17.6L	-AG-A	· T T	A	CAT-TG-	C	GTA TT-GC	TG-A-C	A1	٢G
Gmhs 17.5E	AC T	·T	A	CA	тт	ATG	A -C	A1	٢G

FIG. 5. Comparison of 5' flanking region of soybean HS genes. The nucleotide sequences of Gmhs17.5-M, Gmhs17.6-L, and Gmhs17.5-E (13) are aligned with Gmhs17.5-M as a reference. Numbering is from the transcription start site distal to the TATA motif (heavy line) of Gmhs17.5-M. Broken lines indicate identical bases; spaces are included for maximum alignment. The HS consensus sequences are designated by light underlines. Asterisks denote nucleotide homology to HS core inverted repeat at a minimum of 80%.

					10			20			30				40			50			÷
Gmhsp	17.5E	ATG	TCT	CTG	ÅTT.	CCA	GGT	τṫc	TTC	GGT	CCC	CGA	AGG	AGC	AAC	GTC	TTC	GAT	CCA	TTC	TCA
		Met	Ser	Leu	110	Pro	Gly	Phe	Phe	Gly	Gly	Arg	Arg	Ser	Asn	Val	Phe	Asp	Pro	Phe	Ser C
Gmhap	17.5M						A	A-T				A				0			1		U.
C	17.61			T			Ser A	A-T				-C-				G					
Gmnap	17.0L						Ser	11.				Pro									
HS 6	871						A								-GT	T		C	T		c
							Ser								Ser						
					_						-							110			120
					70			80		CAT			CTT		ÅCT	тст	TCT	GTT	тст	GCT	GAA
Gmhap	17. 5 £	CIC	GAC	AIG	Tro	GAI	Pro	Phe	i vs	Ann	Phe	His	Vai	Pro	Thr	Ser	Ser	Val	Ser	Ald	Glu
Gmbso	17.5M			G									T				CT-				
				Val									Phe				Leu				
Gmhap	17.6L		T																		
												••						C			
HS 6	871		1	G		C		•••				Pro	Phe			A0 -		Leu			
				401																	
					130			140			150				160			170			180
Gmhap	17.5E	AAT	TCT	GCA	ŤΤC	GTG	AGC	AĊT	CGT	GTG	GAŤ	TGG	AAG	GAG	ACC	CCA	GAG	GCA	CAC	GTG	TTC
		Asn	Ser	Ald	Phe	Val	Ser	Thr	Arg	Val	Asp	Trp	Lys	Giu	Thr	Pro	Glu	Ale	His	Val	Phe
Gmhap	17.5M				T		-A-							4							
Carbon	17.6				T		-4-									- 4-					c
Gmnap	17.00						Asn	-								Gin					Leu
HS 6	871		A	G	T			A	A						A		4				
															230			730			240
	17.55		007		ATT	~	000	cic	AAG		CAG	GAA	GTC	AAG	ĠTT	CAG	ATT	GÁA	GAT	GAT	AGG
Gmhap	17.32	AAG	GUI	GAI		Bra	CIV	1.01	ive		Glu	Giu	Val	Lvs	Val	Gin	11.	Glu	Asp	Asp	Arg
Canhan	17 SM	C								G			G		G						
		Glu																			
Gmhsp	17.6L												G								
										-					~ ~	.		^		-00	
HS 6	5871									G			G	G	C-G	G		61.0		-00	Ara
																		••••		•••	
					250			280			27	,			280			290			300
Gmhao	17.5E	GTT	стт	CAG	250 ATT	AGC	GGA	200 GAG	AGG	AAC	275 GT T	GAA	AAG	GAA	200 GAC	AAG	AAC	290 GAC	ACG	TGG	300 CAT
Gmh s p	17.5E	GTT Val	CTT	CAG	25,0 ATT	AGC	GGA Giy	240 GAG GIU	AGG Arg	AAC Asn	27 GTT Val	GAA GIU	AAG Lys	GAA Giu	280 GAC A s p	AAG Lys	AAC Asn	290 GAC A 3 p	ACG Thr	TGG Trp	300 CAT His
Gmh s p Gmhsp	17.5E 17.5M	GTT Val	CTT Leu	GIn	250 ATT 110	AGC Ser	GGA Giy	240 GAG G I u	AGG Arg	AAC Asn	27 GTT Vel C	GAA Glu G	AAG Lys	GAA Glu	200 GAC A s p	AAG Lys	AAC Asn	290 GAC A 3 p	ACG Thr	TGG Trp	300 CAT His
Gmhsp Gmhsp	17.5E 17.5M	GTT Voi	CTT Leu	Gin	250 5 ATT 5 110 A	AGC Ser	GGA Giy	200 GAG GIU	AGG Arg	AAC Asn	27 GTT Val C Leu	GAA GIU G	AAG Lys	GAA Glu	280 GAC Asp	AAG Lys	AAC Asn	290 GAC Asp	ACG Thr	TGG Trp	300 CAT His
Gmhsp Gmhsp Gmhsp	17.5E 17.5M 17.6L	GTT Vol 	CTT Leu	Gin	250 ATT 110 A	AGC	GGA Giy	240 GAG GIU 	AGG Arg 	AAC Asn	27 GTT Val C Leu	GAA Glu G	AAG Lys	GAA Glu 	280 GAC Asp 	AAG Lys	AAC Asn 	290 GAC A s p	ACG Thr	TGG Trp	300 CAT His
Gmhsp Gmhsp Gmhsp HS (17.5E 17.5M 17.6L 6871	GTT Val 	CTT Leu 	Gin	250 ATT 110 A	AGC	GGA G1y 	200 GAG GIU 	AGG Arg	AAC Asn 	27 GTT Val C Leu	GAA Giu G G	AAG Lys 	GAA Glu 	280 GAC Asp 	AAG Lys 	AAC Asn 	290 GAC Asp 	ACG Thr 	TGG Trp 	300 CAT His
Gmhsp Gmhsp Gmhsp HS (17.5E 17.5M 17.6L 6871	GTT Vei	CTT Leu 	Gin	250 ATT 110 	AGC Ser	GGA Gly 	200 GAG GIU 	AGG Arg 	AAC Asn T	277 GTT Val C Leu 	GAA Glu G G	AAG Lys 	GAA Glu 	280 GAC Asp 	AAG Lys 	AAC Asn 	290 GAC A 3 p 	ACG Thr 	TGG Trp 	300 CAT His
Gmhsp Gmhsp Gmhsp HS (17.5E 17.5M 17.6L 6871	GTT Val 	CTT Leu 	Gin	230 ATT 110 A	AGC	GGA Gly 	200 GAG GIU 	AGG Arg 	AAC Asn T	27 GTT Val C Leu	GAA Glu G G	AAG Lys 	GAA Glu 	280 GAC Asp 	AAG Lys 	AAC Asn T	290 GAC Asp 	ACG Thr 	TGG Trp 	300 CAT His
Gmhsp Gmhsp Gmhsp HS (17.5E 17.5M 17.6L 6871	GTT Vel 	CTT Leu 	CAG G1n 	230 ATT 110 A A	AGC Ser	GGA GIy 	200 GAG GIU 320	AGG Arg 	AAC Asn T	277 GTT Val C Leu 	GAA Glu G G	AAG Lys 	GAA Glu 	280 GAC Asp 340	AAG Lys 	AAC Asn T	290 GAC Asp T 390	ACG Thr 	TGG Trp 	300 CAT His 380
Gmhsp Gmhsp Gmhsp HS (Gmhsp	17.5E 17.5M 17.6L 6871 17.5E	GTT Vei CGC	CTT Leu GTG	CAG G1n 	230 ATT 110 A A 310 CGT	AGC	GGA Gly 	200 GAG GIU 320 GGT	AGG Arg 	AAC Asn T TTC	27 GTT Val C Leu S3 ACG	GAA Glu G G 	AAG	GAA Glu TTC	280 GAC Asp 340 AGA	AAG Lys 	AAC Asn 	290 GAC A 3 P 	ACG Thr 	TGG Trp 	300 CAT His 300 AAA
Gmhap Gmhap Gmhap HS (Gmhap	17.5E 17.5M 17.6L 6871 17.5E	GTT Vel CGC Arg	CTT Leu GTG Voi	GAG GIU	250 ATT 110 A A 310 CGT Arg	AGC Ser AGC Ser	GGA Gly AGT Ser	280 GAG GIU 320 GGT GIY	AGG Arg AAG Lys	AAC Asn T TTC Phe	27 GTT Vol C Leu SSX ACG Thr	GAA Glu G G AGA Arg	AAG Lys A AGG Arg	GAA Glu TTC Phe	240 GAC Asp 340 AGA Arg	AAG Lys TTG Leu	AAC Asn T CCCG Pro	290 GAC Asp T 350 GAG GIu	ACG Thr AAT Asn	TGG Trp GCA Ala	300 CAT His 300 AAA Lys
Gmhsp Gmhsp Gmhsp HS (Gmhsp Gmhsp	17.5E 17.5M 17.6L 6871 17.5E 17.5M	GTT Val CGC Arg	CTT Leu GTG Vol	GAG GIU	250 ATT 110 A A 310 CGT Arg	AGC	GGA Gly AGT Ser C	200 GAG GIU 320 GGT GIY C	AGG Arg AAG Lys C Asn	AAC Asn 	ZT GTT Val C Leu S3X ACG Thr -T- Met	GAA Glu G G AGA Arg G	AAG Lys 	GAA Glu TTC Phe	280 GAC Asp 340 AGA Arg G	AAG Lys TTG Leu	AAC Asn 	290 GAC A 3 p T 390 GAG G1 u	ACG Thr AAT Asn	TGG Trp GCA Alg T	300 CAT His 300 AAA Lys
Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L	GTT Val CGC Arg	CTT Leu GTG Vol	CAG Gin Gin Gin Gin Gin	250 ATT 110 A A 310 CGT Arg	AGC Ser AGC Ser	GGA Gly AGT Ser 	200 GAG GIU 320 GGT GIy C	AGG Arg AAG Lys C Asn	AAC Asn 	ZTT GTT Vai C Leu St ACG Thr -T- Met -T-	GAA Glu G G AGA Arg G	AAG Lys 	GAA Glu TTC Phe 	280 GAC Asp 340 AGA Arg G	AAG Lys TTG Leu 	AAC Asn T CCCG Pro A	280 GAC Asp 	ACG Thr AAT Asn 	TGG Trp GCA Alg T	300 CAT His 300 AAA Lys
Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L	GTT Val CGC Arg 	CTT Leu GTG Vol	CAG GI GI GI GI GI GI GI GI GI GI GI GI GI	250 ATT 1 1 0 A 310 CGT Arg 	AGC	GGA Gly AGT Ser C	200 GAG GIU 320 GGT GIY C	AGG Arg AAG Lys C Asn	AAC Aan T TTC Phe 	ZX GTT Vai C Leu SX ACG Thr -T- Met	GAA Glu G G AGA Arg G	AAG Lys A AGG Arg	GAA Glu TTC Phe 	280 GAC Asp 340 AGA Arg G	AAG Lys TTG Leu 	AAC Asn T CCCG Pro A	280 GAC A 3 p T 380 GAG G1u 	ACG Thr AAT Asn 	TGG Trp GCA Ala T	300 CAT His 380 AAA Lys
Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Vel CGC Arg 	CTT Leu GTG Vol 	CAG Gin 	230 ATT 110 A 310 CGT Arg A	AGC Ser AGC Ser 	GGA Gly AGT Ser 	200 GAG GIU 320 GGT GIY C	AGG Arg AAG Lys Aan 	AAC Asn 	ZX GTT Val C Leu ACG Thr -T- Met GT	GAA Giu G G AGA Arg G G	AAG Lys A AGG Arg 	GAA Glu Phe 	280 GAC Asp 340 AGA Arg G	AAG Lys TTG Leu 	AAC Asn T CCCG Pro A	280 GAC A sp T 380 GAG GI u 	ACG Thr AAT Aan 	TGG Trp GCA Alg T T	300 CAT His 300 AAA Lys
Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Val CGC Arg 	CTT Leu GTG Vol 	CAG GIn 	230 ATT 1 1 0 A 310 CGT Arg A	AGC Ser AGC Ser	GGA Gly Ser 	200 GAG GIU 320 GGT GIY C A	AGG Arg AAG Lys C Asn 	AAC Asn 	27 GTT Vai C Leu ACG Thr -T- Met GT- Vai	GAA Glu G G AGA Arg G G	AAG Lys A AGG Arg	GAA Glu Phe T	280 GAC Asp 340 AGA Arg G	AAG Lys TTG Leu 	AAC Asn 	290 GAC Asp T 390 GAG Glu 	ACG Thr AAT Asn 	TGG Trp GCA Ala T T	300 CAT His 300 AAA Lys
Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Val CGC Arg 	CTTT Leu GTG Vol 	GAG GIU GIU GIU GIU	230 ATT 110 	AGC	GGA Gly Ser 	280 GAG G1u 320 GGT G1y C A C	AGG Arg Lys C Asn 	AAC Asn 	277 GTT Vai C Leu S30 ACG Thr -T- Met GT- Vai S30	GAA Glu G G AGA Arg G G	AAG Lys A AGG Arg	GAA Glu Phe 	280 GAC Asp 340 AGA Arg G 	AAG Lys TTG Leu 	AAC Asn 	290 GAC Asp T 330 GAG Glu 	ACG Thr AAT Aan 	TGG Trp GCA Ala T T	300 CAT His 340 AAA Lys 420
Gmhap Gmhap Gmhap HS (Gmhap Gmhap HS (17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Vel CGC Arg 	CTTT Leu GTG Vol 	CAG Gin 	230 ATT 110 	AGC Ser AGC Ser 	GGA Gly Ser 	280 GAG G1u 320 GGT G1y C A C 380 TCT	AGG Arg AAG Lys Asn 	AAC Asn 	27 GTT Vai C Leu S30 ACG Thr -T- Met GT- Val S90 AAT	GAA Glu G G G G G	AAG Lys A AGG Arg 	GAA Glu 	280 GAC Asp 340 AGA Arg 400 ACT	AAG Lys TTG Leu GTC	AAC Asn 	230 GAC A 3 p 	ACG Thr AAT Aan CCT	TGG Trp GCA Alg T T	300 CAT His 330 AAA Lys 420 GAA
Gmhap Gmhap Gmhap HS (Gmhap Gmhap HS (Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E	GTT Val CGC Arg CGC GTG Val	CTTT Leu GTG Vol AATT Asn	CAG Gin 	230 ATT III 	AGC Ser Ser Ser Ser Ser Ser 	GGA Gly Ser GCT Ala	2000 GAG Glu 320 GGT Gly C A C 300 TCT Ser	AGG Arg AAG Lys Asn AtG Met	AAC Asn 	27C GTT Vai C Leu S3C Thr -T- Met GT- Val S3C AAT Aan	GAA Glu G G G G GGG Gly	AAG Lys A AGG Arg 	GAA Glu 	280 GAC Asp 340 AGA Arg 400 ACT Thr	AAG Lys TTG Leu GTC Val	AAC Asn 	280 GAC Asp 	ACG Thr AAT Aan CCT Pro	TGG Trp GCA Ala T T T	300 CAT His 3300 AAA Lys 420 GAA GIu
Gmhap Gmhap Gmhap HS (Gmhap Gmhap HS (Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M	GTT Val CGC Arg GTG Val 	CTTT Leu GTG Vol AATT Asn G-G	CAG GIn 	230 ATT 110 	AGC Ser AGC Ser Ser Ser Ser 	GGA Gly Ser GCT Ala	2000 GAG Glu 3300 GGTy C A C 300 TCT Ser 	AGG Arg Lys Asn Att Met	AAC Asn 	27C GTT Vai C Leu SXG Thr -T Met GT Val Sac AAT Asn 	GAA Glu G G AGA Arg G G GGG Gly T	AAG Lys 	GAAA Glu 	2800 GAC Asp 340 AGA Arg G 400 ACT Thr A	AAG Lys TTG Leu GTC Vol T	AAC Asn 	280 GAC Asp 	ACG Thr AAT Aan CCT Pro C	TGG Trp GCA Ala T T T AAG Ly3 A	300 CAT His 390 AAA Lys GAA GAA GIu
Gmhap Gmhap HS (Gmhap Gmhap HS (Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M	GTT Val CGC Arg GTG Val	GTG Val	GAG GIU GIU GIU GIU GIU GIU GIU GIU GIU GI	230 ATT 110 A 310 CGT Arg A 370 GTG Val 	AGC Ser AGC Ser Ser Ser Ser 	GGA Gly Ser Alg 	2200 GAG Glu 320 GGT Gly C A C 3200 TCT Ser 	AGG Arg Lys Aan Atg Met	AAC Aan 	ZT GTT Val C Leu SX ACG Thr -T- Met GT- Val SQ AAT Asn 	GAA Glu G G G AGA Arg G G GGG Gly T	AAG Lys 	GAA Glu CTC Leu 	2800 GAC Asp 340 AGA Arg G 400 ACT Thr A	AAG Lys TTG Leu GTC Vol T	AAC Asn 	2390 GAC A 3 p 	ACG Thr AAT Aan CCT Pro C	TGG Trp GCA Ala T T AAG Lys A	300 CAT His 300 AAA Lys GAA Glu
Gmhap Gmhap Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L	GTT Vel CGC Arg GTG Vel 	GTG Val	GAG GIU 	230 ATT 110 A 310 CGT Arg A 370 GTG Val A	AGC Ser AGC Ser Ser AGC Ser 	GGA Gly Ser 	2200 GAG Glu 320 GGT Gly C A C 320 TCT Ser 	AGG Arg AAG Lys C Asn ATG Met	AAC Asn 	27 GTT Val C Leu Met GT- Val Sec GT- Val Sec AAT Aan	GAA Glu G G G G G GGG Gly T	AAG Lys 	GAA Glu CTC Leu 	280 GAC Asp AGA Arg G ACT Thr A	AAG Lys TTG Leu GTC Vol T	AAC Asn 	230 GAC Asp 	ACG Thr AAT Aan CCT Pro C A	TGG Trp GCA Ala T T T AAG Lys A	300 CAT His 300 AAA Lys GAA Glu CIu
Gmhap Gmhap HS (Gmhap Gmhap HS (Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 5871	GTT Val CGCC Arg GTG Val 	GTT Leu GTG Vei Asn G-G Glu G-G G-G	CAG GIU 	230 3 ATT 1 1 - A 310 310 310 310 370 370 GTG GTG 4 ATT A 370 A	AGC Ser AGC Ser AGC Ser AGC Ser AGC Ser	GGA Gly AGT Ser GCT Ala 	2800 GAG GIU 3200 GGT GIU 	AGG Arg AAG Lys Aan At G Met 	AAC Aan 	27, GTT Val C Leu Sig ACG Thr -T- Met GT- Val 30, GT- Val 30, GT- Val 	GAA Glu G G G G GGC Gly T 	AAG Lys 	GAA Glu 	280 GAC Asp AGA Arg 	AAG Lys TTG Leu GTC Vol T T	AACC Aan T CCCG Pro A A ACT Thr C 	280 GAC A3p 	ACG Thr AAT Aan CCT Pro C A	TGG Trp GCA Alo T T T AAG Lys 	100 CAT His 300 AAA Lys GAA Glu CIu
Gmhap Gmhap HS (Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5L 6871	GTT Val CGC Arg GTG Val 	GTT Leu GTG Vei GTG Vei Asn G-G Giu G-G GLU G-C	CAG GIU 	230 GATI 110 110 110 110 110 110 110 11	AGC Ser 	GGA Gly Ser GCT Ala	2800 GAG GIU 3200 GGT GIU C C C C C C C 	AGG Arg Lys Asn AtG Met 	AAC Aan 	27, GTT Val C Leu Sig ACG Chr -T- Met GT- Val 30, GT- Val 	GAA Glu G G G G GGC Gly T 	AAGG A AGG Arg GTT Vol 	GAA Glu 	280 GAC Asp 340 AGA Arg G ACT Thr A	AAG Lys TTG Leu GTC Vol T T A	AACC A an 	280 GAC A3p 	ACG Thr AAT Asn CCT Pro C A	TGG Trp GCA Ala T T T AAG Lys A	100 CAT His 300 AAA Lys GAA Glu
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 17.5E 17.5M 17.6L 5871	GTT Val CGC Arg GTG Val 	CTT Leu GTG Vol CTG Vol CTG Vol CTG Vol CTG Vol CTG CG CG CG CG CG CG CG CG CG CG CG CG CG	GAG Glu Glu Glu Glu Glu Glu Glu Glu Glu Gl	230 310 310 CGT 4 Arg A 310 CGT 4 Arg A 370 5 GTG Val A A	AGC Ser AGC Ser AGC Ser 	GGA Gly Ser Alg Alg	280 GAG GIU 320 GGT GIY C A C TCT Ser C Cys C	AGG Arg Lys Aan AtG Met	AAC Aan T TTC Phe Leu GAA Glu 	27, T GTT Vai C Leu Sac GThr -T- Met GT- Vai Sac AAT Asn 	GAA Glu G G G G GGG Gly T T	AAG Lys A AGG Arg GTT Va1 	GAA Glu Phe CTC Leu 	280 GAC Asp 340 AGA Arg G 400 ACT Thr A	AAG Lys TTG Leu GTC Vol T T A	AACC Asn T CCCG Pro A A ACT Thr C 	280 GAC Asp T 330 GAG GIU 410 GTT Val Lie 	ACG Thr AAT Aan CCT Pro C A	TGG Trp GCA Ala T T T AAG Lys A	300 CAT His 300 AAA Lys GAI GAI
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5L 6871	GTT Val CGC Arg GTG Val 	GTG Vol AAT G-G Glu G-G Glu G-C Asp	GAG Glu GAG Glu GLU GLU GLU GLU GLU GLU GLU GLU GLU GLU	230 3 ATT 1 1 e A 310 310 CGT CGT GTG GTG Val A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 A 370 	AGC Ser AGC Ser AGC Ser 	GGA Gly GGT Ala 	280 GAG GIU 320 GGT GGT GGT 	AGG Arg AAG Lys Asn ATG Met 	AAC Aan 	27, T GTT Vai C Leu Sac GT Vai Sec GT Vai Sec AAT Asn 	GAA Glu G G G G GGG Gly T 	AAG Lys 	GAA Glu 	280 GAC Asp AGA Arg ACT Thr A +400 ACT Thr 	AAG Lys TTG Leu GTC Vol T T 	AACC A	280 GAC Asp 	ACG Thr AAT Aan CCT Pro C A	TGG Trp GCA Alg T AAG Lys 	хор САТ Ніз Зар ААА Lys GAA Glu
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap HS (17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Val CCCC Arg GTC Val GTC Val	CTTT Leu GTG Vol Aan G-G Glu G-G Glu G-C Aap	GAG GIU GAG GIU GIU C GIU C GIU C GIU C GIU	230 ATT 110 A 310 CGT CGT CGT CGT CGT CGT CGT CGT	AGC Ser AGC Ser AGC Ser 	GGA Gly GGT GGT	2800 GAG GIU 320 GGT GIY C A C 340 TCT Ser 	AGG Arg AAG Lys Asn ATG Met AAG	AACC Aan 	27C GTT C Leu SSC ACG Thr -T Wet GT Val SRC AAT Asn 4SC AT	GAA GIU G G AGA Arg G GGG GIy T GGAA	AAGG 	GAA Glu 	200 GAC Asp Aga 	AAG Lys TTG Leu GTC Vol T T A	AACC Aan 	280 GAC Asp 	ACG Thr AAT Asn CCT Pro C A	TGG Trp GCA Ala T T T AAG Lys 	300 CAT His 300 AAA Lys GAA GIu
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.6L 6871 17.5E 17.5E	GTT CGCC Arg GTG Val GAG Glu	GTG GTG Voi GTG GIU G-G GIU G-C Asp GTT Voi	GAG Glu 	230 3 ATT 1 1 0 	AGCC Ser AGCC Ser AGCC Ser Cor CCT Pro	GGA GIY GIY GIY GIY GIY GIY GGT Alg GAT Asp	2800 GAG GIU 320 GGT GIY 	AGG Arg AAG Lys Aan Aan Met AAG Lys	AACC Aan 	27, T GTT C Leu S3C ACG Thr -T- S4C GT Val S4C GT Val S4C GT Leu S4C GT Leu S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT Thr -T S4C GTT - S4C GTT - S4C - S4C GTT - S4C GTT - S4C - - S4C - - S4C - - - - - - - - - - - - - - - - - - -	GAA GIU G G G G GGG GIY T GGAA GIU	AAGG Lys 	GAA Glu TTC Phe T CTC Leu TCT Ser	200 GAC Asp Aga Acr 	AAG Lys TTG Leu GTC Vol T T T A	AACC Pro 	2300 GAC Asp 	ACG Thr AAT Asn CCT Pro C A	TGG Trp GCA Ala T T AAG Lys 	300 CAT His 300 AAA Lys 420 GAA GAA
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap HS (Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.6L 5871 17.5E 17.5E	GTT CGCC Arg GTG Val GAG Glu	GTG Vai	GAG Glu 	230 ATT A A 310 CGT A 370 A 370 A 370 A 370 A 370 A 370 A 370 A A A A A A A 	AGC Ser AGC Ser AGC Ser 	GGA GIy GIy GIY GGT AGT AGT AIG AGT ASP	280 GAG GIU 320 GGT GIY 	AGG Arg Lys Aan ATG Met AAG Lys 	AAC Asn 	27,5 GTT C Leu Sign ACG Chr -T Met -T Val 38,6 AAT Asn 48,6 ATT 11e 	GAA GIU G G G GGU T G GGGG GIY T 	AAGG Lys 	GAA Glu CTC Leu CTC Leu CTC Leu 	280 GAC Asp -	AAG Lys TTG Leu GTC Vol T T A	AAC Asn 	2300 GAC Asp GAG Glu GT GT Val Ite 	ACG Thr AAT Aan CCT Pro C A	TGG Trp GCA Ala T 	300 CAT His 390 AAA Lys 420 GAA
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap Gmhap HS (Smhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.6L 6871 17.5E 17.5M 17.5E	GTT Val GTG Val GTG GTG GAG Glu 	CTT Leu GTG Vol AAT G-G Glu G-G Glu G-C Glu G-C Glu C-C C T Val	GAG Glu 	230 ATT 110 A 310 CGT A 370 GTG GTG A 370 	AGC Ser AGC Ser 	GGA GI GGI GI GGI GGI GGI GGI GGI Alo C C GAT Asp C C C C C C C C C C C C C C C C C C C	2400 GAG GIU 3300 GG1y C 	AGG Arg Lys Asn ATG Met ATG Lys AAG	AAC Aan 	ZTT GTT C Leu Star ACG Thr -T Met GT Yal 30 AAT 40 AAT 40 C C 40 C C 	GAA GIU G G G G GIU G GGG GIY G GGGG GIY T 	AAGG Lys 	GAA Glu 	220 GAC Asp ACT Thr ACT Thr 	AAG Lys TTG Leu GTC Vol T	AAC Asn 	250 GAC Asp GTT Val Ite 	ACG Thr AAT Aan CCT Pro C A 	TGC Trp GCA Aia T 	300 CAT His 300 AAA Lys GAA Glu CAA
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap HS (Cmhap HS (Cmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.5E 17.5M 17.5E	GTT Val GTG GTG GIG GIG CAG GIU	GTG Vol AAT G-G Glu G-G Glu G-C Asp GTT Val	GAAG Glu 	230 ATT 110 A 310 CGT A 370 GTG GTG A 370 A A 370 A 370 A 370 A 370 A 370 A 370 A 370 	AGC Ser AGC Ser AGC Ser AGC Ser CCT Pro T Ser	GGA GI y GI y GI y GI y GI y GI y GI y GI y	2000 GAG GIU 320 GGT GIY C 	AGG Arg AAG Lys Aan ATG Met AAG Lys AAG	AAC Aan 	27,7 GTT C Leu SacG Thr -T- Met GT- Val SacG GT- Val SacG AAT 400 AAT 400 AT 400 	GAA GIU G G G G GGG GIY G GGG GIY T 	AAG Lys 	GAA Glu 	280 GAC Asp Aga Aga Arg AcT Thr 	AAG Lys GTC Vol 	AACC Ain ACT Thr ACT ACT ACT 	2300 GAC Asp 410 GTT Vol Lie 	ACG Thr AAT A3n CCT Pro C A	TGG Trp GCA Alg T T T AAG Lys 	300 CAT His 300 AAA Lys 420 GAA Glu
Gmhap Gmhap HS (Gmhap Gmhap Gmhap Gmhap HS (Gmhap Gmhap Gmhap	17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871 17.5E 17.5M 17.6L 6871	GTT Vol CGCC Arg GTG Vol GAG Glu 	GTG GTG Vol AAT G-G Glu G-C Asp GTT Val A	GAA GAU GAU GAU GAU GAU GAU C GIN C GIN C GIN	230 ATT 110 A 310 CGT Arg A 370 GTG Val A 	AGC Ser 	GGA AGT GGT Аlg GAT Азр 	240 GAG GIU 320 GGT GIY C GGT TCT Ser C 	AGG Arg AAG Lys Asn ATG Met AAG Lys 	AAC Aan 	27 27 27 27 2 2 2 2 2 2 2 2 2 2 2 2 2	GAA GIU G GIU G GGG GIY GGAA GIU GGAA GIU 	AAG Lys A AGG Arg GTT Vol ATC lie 	GAA Glu 	250 GAC Asp 340 AGA Arg G 460 ACT Thr A GGT GI y 	AAG Lys TTG Leu GTC Vol T T A TGA	AACC Asn 	2300 GAC A3p 	ACC Thr AAT A3n CCT Pro C A	TGG Trp 	300 CAT His 300 AAA Lys GAA GIU

FIG. 6. Nucleotide sequence and deduced amino acid sequence comparisons of four soybean low-molecular-weight HS proteins. The sequence of *Gmhsp17.5-E* is used as reference. Dashed lines indicate identical nucleotides, whereas nucleotide changes are indicated by appropriate letters. Amino acid differences are listed below corresponding codons. Blank spaces in *Gmhsp17.5-M* and *HS6871* are predicted deletions which are added to maintain maximum homology. Sequence *Gmhsp17.5-E* is from Czarnecka et al. (13), and *HS6871* is from Schöffl et al. (46).

a diverse group of small HS proteins. The soybean and *Drosophila* (49) plots were constructed from the deduced sequences of completely sequenced genomic clones, whereas the *Xenopus* (7) and *C. elegans* (42) plots are from published amino acid sequences derived from the DNA

sequences of partial-length cDNA clones. The most prominent common feature of these profiles is the major hydrophilic peak centered around amino acid residue 95, which is flanked by a carboxy-proximal hydrophobic region. Although the hydropathic nature of this region seems to be



FIG. 7. Deduced amino acid sequences from the carboxy halves of 10 HS proteins. The sequences have been aligned to maximize homology with the four *Drosophila* small HS proteins (18). Numbers under the sequences correspond to the amino acid position in the composite soybean protein. The organisms of origin and molecular weights of the proteins are given at the carboxy termini. The sequences of *C. elegans* are a composite of two cDNAs (42), and the *Xenopus* sequences represent cDNAs X4 and X5 (7). Boxes indicated homology with the composite amino acid sequence of four soybean HS genes.

important, the actual amino acids used in this region showed considerable diversity among the organisms studied. This type of analysis also suggests a closer structural relationship between soybean and *Drosophila* proteins than is evident by a simple analysis of amino acid homology and suggests that the soybean proteins have more in common with *Drosophila* small HS proteins than with either X. laevis or C. elegans.

DISCUSSION

HS genes as multigene families. The relatively high abundance and distinctly lower molecular weight of the more than 20 15- to 18-kDa HS proteins of soybean initially suggested that perhaps these HS proteins could be unique to plants and perhaps regulated in a manner different from those of animals. Analysis of data presented here, however, together with other work (12, 23, 44, 45), demonstrates that the low-molecular-weight HS proteins of soybean represent multigene families analogous to the small HS groups first characterized in D. melanogaster (18). The largest soybean family, consisting of thirteen proteins, is partially described here and is represented by genomic clones pE2019, pL2005, pM2005, and pHS6871 and cDNA clones pFS2019, pFS2005, pFS1968, and pCE53 (13, 44-46). Additional families of small HS proteins include cDNA clone pCE75 (15-kDa class), pFS2033 (21- to 24-kDa class), and possibly pCE54 (27-kDa class) (13, 23, 44, 46). Other groups of small HS proteins are known to exist, including HS proteins which are transported into the chloroplasts (E. Vierling, unpublished data).

Coding region analysis. A comparison of the predicted amino acid sequences reveals approximately 90% homology among the soybean HS proteins analyzed to date. A majority of the nucleotide changes represent silent substitutions. A comparison of the base changes with regard to composition and location suggests a duplication mechanism for the origin of these HS genes. Likewise, the close linkage of the four low-molecular-weight HS genes in D. melanogaster at locus 67B suggests that these have a common origin and arose by gene duplication and inversion events (49). The very high homology between the coding regions of the low-molecularweight HS genes of soybean, as well as multiple-cross homologies in the 5' flanking regions, also suggests a common origin by duplication. However, regardless of origin, the maintenance of these multiple homologies may represent a regulatory refinement for the assurance of high expression of these genes under specific physiological conditions, e.g., HS.

Comparison with the *Drosophila* small HS proteins based simply on linear amino acid composition shows low but significant homology to the soybean HS proteins. However, a more general structural relationship is revealed by hydropathy analyses (27). The major hydrophilic peak, which is characteristic of the small HS proteins of *D. melanogaster* (49), is clearly present in the soybean proteins



FIG. 8. Hydropathic profiles of deduced amino acid sequences of (A) soybean Gmhsp17.6-L, (B) Drosophila hsp27 (49), (C) Xeno-pus hsp30 (7), and (D) C. elegans hsp16 (42). Plots were constructed by the method of Kyte and Doolittle (27) by progressively moving along the amino acids sequence and averaging the hydropathy index for nine amino acids. Points above the horizontal line correspond to hydrophobic regions, and points below this line are hydrophilic. The plots are aligned along the major hydrophilic peak which is characteristic of the small HS protein.

from position 95 to 105 (Fig. 8). A total of four common domains of similarity have been identified (13). A striking conservation of Asp-Gly-Val-Leu-Thr occurs in a very hydrophobic domain (positions 125 to 140; Fig. 8) in all the *Drosophila* small HS proteins, α -crystallin, and the soybean genes, except that in soybean an Asn replaces Asp. The overall similarities in the hydropathy profiles and a significant level of amino acid homology are suggestive of the conservation of functional domains among this broad group of proteins from highly diverged groups of organisms.

The selective localization of low-molecular-weight HS proteins in soybean to organellar fractions (e.g., nuclei, mitochondria, and ribosomes) during HS is likely the basis of thermoprotection (22). Selective localization is also known for some low-molecular-weight HS proteins of *Drosophila* (2, 3). It has been suggested (28, 46, 49) that the similarities in the protein domains derived from hydropathy analysis between the small HS proteins and the bovine α -crystallin lens protein may be related to common aggregation properties and possibly other types of protein-protein interactions involved in function and localization.

HS promoter analysis. In addition to a TATA-like motif characteristic of many eucaryotic genes, soybean HS genes also show a great deal of similarity in the DNA sequence composition of the 5' flanking region of HS genes of D. *melanogaster* and X. *laevis*. The best-characterized sequence element is the HS consensus core of D. MOL. CELL. BIOL.

melanogaster, CT-GAA--TTC-AG (38). In heterologous systems, in high copy number, this sequence has been shown to be sufficient for thermal activation of transcription when placed 13 or 19 nucleotides upstream to the TATA-box (39). Although the Drosophila HS consensus functions inductively in heterospecific systems, the temperature optimum for expression corresponds to the recipient cells (9, 11, 32, 33). In most cases, an 8 of 10 match with the HS consensus core is required for thermal induction. Exceptions to this rule suggest that a 6 of 10 match can function when additional copies of the consensus are present (1, 8, 39). In all four of the soybean genes discussed here, the TATAproximal HS consensus element overlaps with a second HS consensus sequence with lower homology. Additional HS consensus elements with 8 of 10 homology match are identified further upstream in the soybean genes. These redundant elements are located at positions -169 to -182 nucleotides in pL; -499 to -512 nucleotides in pM; -358 to -371nucleotides in pE; and -212 to -225, -221 to -234, -232 to -245, and -263 to -276 nucleotides in HS6871. This finding is not unexpected, since most HS genes in other organisms also contain multiple copies of the HS consensus core sequences located within 250 nucleotides of the start of transcription (8).

Two protein-binding sites are implicated in the activation of HS genes (53), and a specific activating protein has been described that binds the HS consensus region of hsp82 gene chromatin in vitro (54). A specific HS transcription factor and a general transcription factor (the A factor) are required for active in vitro transcription of the hsp70 gene by RNA polymerase II (37). The presence of redundant HS consensus sequences in soybean HS genes suggests the possibility of cooperative binding of multiple HS transcription factor proteins either to effect high levels of transcription or to modulate the level under somewhat different physiological states relative to stress or both. It has been proposed (8, 16) that redundant HS consensus core elements are required for efficient thermal induction when the gene is in low copy number and must compete with other HS genes for transcriptional factors or when a suboptimal promoter configuration exists because of poor spacing or homology.

A decameric palindromic sequence in D. melanogaster is found 5' distal to the HS consensus element in similar positions in hsp26 and hsp70 and in slightly modified form in hsp22 (49). It is intriguing to note that this decameric palidromic sequence is partially homologous to the HS core inverted repeat (AGAAATTTCT; CTnGAAnnTTCnAG). An analysis for homology to this decameric palindromic sequence showed 70 to 80% homology in the soybean HS genes in numerous locations. Interestingly, in the pL clone, 70% homology to this decameric palindromic sequence is located four times in the sequence 5' to the TTTAAATA motif. In each case, this sequence is within or partially within a sequence of at least 60% homology to an HS consensus sequence (centered at -35, -94, -144, and -154 bp from the 5' end of the TTTAAATA motif). The 5' end of the sequence centered at -94 is 28 bp upstream from a second potential TATA-like sequence of TAAAATA. A similar analysis of the pM clone 5' sequence shows 80% homology with the Drosophila decameric palindrome at -88, -329, -350, and -621 bp from the TTTAAATA motif. There are numerous matches of 70% homology, but none of these are located within or adjacent to an HS consensus sequence. This element is also located in the upstream region of Gmhsp17.5-E (-186 and -354) and HS6871 (-35, -87, -138, and -310). The element at -310 in HS6871 is 40 bp upstream from a TTTAAATA motif. Although this decameric palindromic sequence has not been shown to be functionally significant in the thermal activation of transcription in animal systems, the presence of a similar sequence in soybean HS genes suggests evolutionary conservation of this element and therefore a possible role in modulating HS gene expression.

In *D. melanogaster*, numerous other stress treatments are effective in inducing HS proteins (4). Most of these other stresses are not effective in soybean; however, arsenite and, to a lesser extent, cadmium induce $poly(A^+)$ RNAs homologous to some HS cDNAs (12). In many organisms, heavy metals, such as cadmium, induce the synthesis of metalbinding proteins, the metallothioneins, for which specific regulatory elements have been identified in the 5' flanking regions of these genes (19). Sequence homology searches for these elements located four regions of 75% homology, for example, in the 5' flanking sequence of pM. Several homologies of greater than 70% were located within other soybean HS genes. Since some HS genes in soybean are induced by heavy metals (12), these homologies are suggestive of possible functional significance.

In the soybean HS genes, slight homology to the canonical sequence 5'-GG^C_CCAATCT'-3' or CCAAT box (6) is centered 53 and 50 bp upstream from the 5' end of the TT-TAAATA motif of pE and pM. Higher homology is located 145 bp upstream from the TTTAAATA motif in HS6871 (46), but no corresponding homology is found in pL. The CCAAT box homology present in some of the soybean genes may only be coincidental, since the Drosophila low-molecular-weight HS-protein genes show no obvious homology to the CCAAT box sequence (49).

DNA sequence analysis of the low-molecular-weight HS genes of soybean supports the view that the molecular mechanisms involved in the thermal induction of HS genes are highly conserved among eucaryotes. In addition to the HS consensus core, other DNA sequences, such as the metal response element, simian virus 40 enhancer core, and potential Z-DNA stretches have been identified in the 5' flanking regions. The presence of these homologies suggests a general conservation in DNA sequences that control transcription in eucaryotes and raises the possibility that these soybean HS genes may be subject to a variety of controls in addition to HS. The continued structural analyses of HS genes and HS proteins coupled with in vitro mutagenesis transformation or expression experiments should provide a better basis for understanding the HS genes of plants.

ACKNOWLEDGMENTS

We thank Jerry Slightom and Yu Ma for providing the λ_{1059} genomic library. We acknowledge the technical assistance of Kenlock Westberry III. We thank Burlyn Michel for writing and modifying computer plotting programs and Joyce Kochert for graphics assistance. We thank Janice Kimpel, Elizabeth Vierling, and John Walker for their helpful comments in preparing the manuscript.

This work was supported by a research contract from Agrigenetics Research Associates, Limited.

LITERATURE CITED

- 1. Ayme, A., R. Southgate, and A. Tissieres. 1985. Nucleotide sequences responsible for the thermal inducibility of the *Drosophila* small heat-shock protein genes in monkey COS cells. J. Mol. Biol. 182:469–475.
- Arrigo, A. P., and C. Ahmad-Zadeh. 1981. Immunofluorescence localization of a small heat shock protein (hsp23) in salivary gland cells of *Drosophila melanogaster*. Mol. Gen. Genet.

184:73-79.

- Arrigo, A. P., S. Fakan, and A. Tissieres. 1980. Localization of the heat shock-induced proteins in *Drosophila melanogaster* tissue culture cells. Dev. Biol. 78:86-103.
- 4. Ashburner, M., and J. J. Bonner. 1979. The induction of gene activity in *Drosophila* by heat shock. Cell 17:241-254.
- Barker, R. F., K. B. Idler, D. V. Thompson, and J. D. Kemp. 1983. Nucleotide sequence of the T-DNA region from the *Agrobacterium tumefaciens* octopine Ti plasmid pTi 15955. Plant Mol. Biol. 2:335–350.
- Benoist, C., K. O'Hare, R. Breathnach, and P. Chambon. 1980. The ovalbumin gene-sequence of putative control regions. Nucleic Acids Res. 8:127–142.
- 7. Bienz, M. 1984. Developmental control of the heat shock response in *Xenopus*. Proc. Natl. Acad. Sci. USA 81:3138-3142.
- 8. Bienz, M. 1985. Transient and developmental activation of heat-shock genes. Trends Biochem. Sci. 10:157–161.
- Bienz, M., and H. R. B. Pelham. 1982. Expression of a Drosophila heat-shock protein in Xenopus oocytes: conserved and divergent regulatory signals. EMBO J. 1:1583–1588.
- Blattner, F. R., A. Blechl, K. Denniston-Thompson, H. E. Faber, J. E. Richards, J. E. Slightom, P. W. Tucker, and O. Smithies. 1978. Cloning human fetal γ globin and mouse α-type globin DNA: preparation and screening of shotgun collections. Science 202:1279–1284.
- Corces, V., A. Pellicer, R. Axel, and M. Meselson. 1981. Integration, transcription and control of a *Drosophila* heat shock gene in mouse cells. Proc. Natl. Acad. Sci. USA 78:7038–7042.
- Czarnecka, E., L. Edelman, F. Schöffl, and J. L. Key. 1984. Comparative analysis of physical stress responses in soybean seedlings using cloned heat shock cDNAs. Plant Mol. Biol. 3:45-58.
- Czarnecka, E., W. B. Gurley, R. T. Nagao, L. Mosquera, and J. L. Key. 1985. DNA sequence and transcript mapping of a soybean gene encoding a small heat shock protein. Proc. Natl. Acad. Sci. USA 82:3726-3730.
- 14. DiDomenico, B. J., G. E. Bugaisky, and S. Lindquist. 1982. Heat shock and recovery are mediated by different translational mechanisms. Proc. Natl. Acad. Sci. USA 79:6181–6185.
- 15. DiDomenico, B. J., G. E. Bugaisky, and S. Lindquist. 1982. The heat shock response is self-regulated at both the transcriptional and posttranscriptional levels. Cell 31:593–603.
- Dudler, R., and A. A. Travers. 1984. Upstream elements necessary for optimal function of the hsp70 promoter in transformed flies. Cell 38:391–398.
- Favaloro, J., R. Treisman, and R. Kamen. 1980. Transcription maps of polyoma virus-specific RNA: analysis by twodimensional nuclease S1 gel mapping. Methods Enzymol. 65:718-749.
- 18. Ingolia, T. D., and E. A. Craig. 1982. Four small *Drosophila* heat shock proteins are related to each other and to mammalian α -crystallin. Proc. Natl. Acad. Sci. USA 79:2360-2364.
- Karin, M., A. Haslinger, H. Holtgreve, R. I. Richards, P. Krauter, H. M. Westphal, and M. Beato. 1984. Characterization of DNA sequences through which cadmium and glucorticoid hormones induce human metallothionein-II_A gene. Nature (London) 308:513-519.
- 20. Karn, J., S. Brenner, L. Barnett, and G. Cesareni. 1980. Novel bacteriophage λ cloning vector. Proc. Natl. Acad. Sci. USA 77:5172-5176.
- Kelley, P. M., and M. J. Schlesinger. 1982. Antibodies to two major chicken heat shock proteins cross-react with similar proteins in widely divergent species. Mol. Cell. Biol. 2:267-274.
- 22. Key, J. L., E. Czarnecka, C. Y. Lin, J. Kimpel, C. Mothershed, and F. Schöffl. 1983. A comparative analysis of the heat shock response in crop plants, p. 107–118. *In* D. D. Randall, D. G. Blevins, R. L. Larson, and B. J. Rapp. (ed.), Current topics in plant biochemistry and physiology, vol. 2. University of Missouri, Columbia, Mo.
- Key, J. L., W. B. Gurley, R. T. Nagao, E. Czarnecka, and M. A. Mansfield. 1985. Multigene families of soybean heat shock proteins. *In L. van Vloten-Doting, G. Groot, and T. Hall (ed.)*,

NATO Advanced Studies Institute molecular form and function simplex virus

- of the plant genome. Plenum Publishing Corp., New York.
 24. Key, J. L., J. A. Kimpel, C. Y. Lin, R. T. Nagao, E. Vierling, E. Czarnecka, W. B. Gurley, J. K. Roberts, M. A. Mansfield, and L. Edelman. 1985. The heat shock response in soybean, p. 161–179. In J. L. Key and T. Kosuge (ed.), Cellular and molecular biology of plant stress, vol. 22. Alan R. Liss, Inc., New York.
- Key, J. L., J. A. Kimpel, E. Vierling, C. Y. Lin, R. T. Nagao, E. Czarnecka, and F. Schöffl. 1985. Physiological and molecular analyses of the heat shock response in plants, p. 327–348. In B. G. Atkinson and D. B. Walden (ed.), Changes in eukaryotic gene expression in response to environmental stress. Academic Press, Inc., New York.
- Key, J. L., C. Y. Lin, and Y. M. Chen. 1981. Heat shock proteins of higher plants. Proc. Natl. Acad. Sci. USA 78:3526-3530.
- Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105-132.
- Lin, C. Y., J. R. Roberts, and J. L. Key. 1984. Acquisition of thermotolerance in soybean seedlings. Plant Physiol. 74:152-160.
- Linquist, S. 1980. Varying patterns of protein synthesis in Drosophila during heat shock: implications for regulation. Dev. Biol. 77:463–479.
- 30. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- 31. Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. 65:449–560.
- McMahon, A. P., T. J. Novak, R. J. Britten, and E. H. Davidson. 1984. Inducible expression of a cloned heat shock fusion gene in sea urchin embryos. Proc. Natl. Acad. Sci. USA 81:7490-7494.
- Mirault, M. E., R. Southgate, and E. Delwart. 1982. Regulation of heat-shock genes: a DNA sequence upstream of *Drosophila* hsp70 genes is essential for their induction in monkey cells. EMBO J. 1:1279-1285.
- Nagao, R. T., D. M. Shah, V. K. Eckenrode, and R. B. Meagher. 1981. Multigene family of actin-related sequences isolated from a soybean genomic library. DNA 1:1–9.
- 35. Nordheim, A., and A. Rich. 1983. Negatively supercoiled simian virus 40 DNA contains Z-DNA segments within transcriptional enhancer sequences. Nature (London) 303:674–678.
- 36. O'Connor, D., and J. T. Lis. 1981. Two closely linked transcription units within the 65B heat shock puff locus of D. melanogaster display strikingly different regulation. Nucleic Acids Res. 9:5075-5092.
- Parker, C. S., and J. Topol. 1984. A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene. Cell 37:273-283.
- Pelham, H. R. B. 1982: A regulatory upstream promoter element in the *Drosophila* hsp70 heat-shock gene. Cell 30:517–528.
- 39. Pelham, H. R. B., and M. Bienz. 1982. A synthetic heat-shock promoter element confers heat-inducibility on the herpes

simplex virus thymidine kinase gene. EMBO J. 1:1473-1477.

- Proudfoot, N. J., and G. G. Brownlee. 1976. 3' Non-coding region sequences in eukaryotic messenger RNA. Nature (London) 263:211-214.
- 41. Pustell, J., and F. C. Kafatos. 1982. A convenient and adaptable package of DNA sequence analysis programs. Nucleic Acids Res. 10:51–59.
- 42. Russnak, R. H., D. Jones, and E. P. M. Candido. 1983. Cloning and analysis of cDNA sequences coding for two 16 kilodalton heat shock proteins (hsps) in *Caenorhabditis elegans*: homology with the small hsps of *Drosophila*. Nucleic Acids Res. 11:3187-3205.
- 43. Schlesinger, M. J., M. Ashburner, and A. Tissieres. 1982. Heat shock: from bacteria to man. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- 44. Schöffl, F., and J. L. Key. 1982. An analysis of mRNAs for a group of heat shock proteins of soybean using cloned cDNAs. J. Mol. Appl. Genet. 1:301–314.
- 45. Schöffl, F., and J. L. Key. 1983. Identification of a multigene family for small heat shock proteins in soybean and physical characterization of one individual gene coding region. Plant Mol. Biol. 2:269-278.
- 46. Schöffl, F., E. Raschke, and R. T. Nagao. 1984. The DNA sequence analysis of soybean heat-shock genes and identification of possible regulatory promoter elements. EMBO J. 3:2491-2497.
- 47. Shah, D. M., D. E. Rochester, G. G. Krivi, C. M. Hironaka, T. J. Mozer, R. T. Fraley, and D. C. Tiemeier. 1985. Structure and expression of maize hsp 70 gene, p. 181–200. *In* J. L. Key and T. Kosuge (ed.), Cellular and molecular biology of plant stress, vol. 22. Alan R. Liss, Inc., New York.
- 48. Slightom, J. L., A. E. Blechl, and O. Smithies. 1980. Human fetal G_{γ} and A_{γ} -globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638.
- 49. Southgate, R., A. Ayme, and R. Voellmy. 1983. Nucleotide sequence analysis of the *Drosophila* small heat shock gene cluster at locus 67B. J. Mol. Biol. 165:35-57.
- 50. Southgate, R., M.-E. Mirault, A. Ayme, and A. Tissieres. 1985. Organization, sequences, and induction of heat shock genes, p. 3-30. In B. G. Atkinson and D. B. Walden (ed.), Changes in eukaryotic gene expression in response to environmental stress. Academic Press, Inc., Orlando, Fla.
- Vieira, J., and J. Messing. 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19:259–268.
- 52. Vierling, E., and J. L. Key. 1985. Ribulose 1,5-bisphosphate carboxylase synthesis during heat shock. Plant Physiol. 78:155-162.
- 53. Wu, C. 1984. Two protein-binding sites in chromation implicated in the activation of heat-shock genes. Nature (London) 309:229-234.
- 54. Wu, C. 1984. Activating protein factors binds *in vitro* to upstream control sequences in heat-shock gene chromation. Nature (London) 311:81–84.