Four small *Drosophila* heat shock proteins are related to each other and to mammalian α -crystallin

(DNA sequence/sequence comparisons/gene family)

THOMAS D. INGOLIA AND ELIZABETH A. CRAIG

Department of Physiological Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706

Communicated by Daniel E. Koshland, Jr., January 18, 1982

ABSTRACT The primary base sequence of the protein coding regions of the four small heat shock genes of Drosophila melanogaster present at cytological locus 67B has been determined. A single open reading frame large enough to encode a small heat shock protein is found for each gene. The molecular weights of the predicted proteins are in good agreement with experimentally determined values obtained from gel electrophoresis. The predicted amino acid sequences of the four small heat shock genes show striking homologies over \approx 50% of their lengths. This region of extensive homology extends from about amino acid 85 to amino acid 195 out of a total of ≈200 amino acids. Comparison of the predicted sequence with the known sequences of other proteins revealed a remarkable similarity between this region of homology and the corresponding region of mammalian α -crystallin. The possible functional significance of this structural similarity is discussed.

Heat shock and a variety of other stimuli result in a dramatic change in the pattern of gene expression in Drosophila melanogaster (for review, see ref. 1). Transcription of most genes is suppressed and expression of a small set of previously inactive genes is enhanced. A similar response is observed in cells of a wide variety of species including yeast (2, 3), Dictyostelium (4), soy bean (5), and chicken and man (6). Little is known about the physiological role of the heat shock response or the induced proteins. However, the heat shock response has been shown to provide protection from thermal killing in Drosophila (7), yeast (8), and Dictyostelium (4). Cytological and cell fractionation studies indicate that many of the Drosophila heat shock polypeptides reside in the nucleus (9-11) associated with either the chromosomes (9) or the nucleoskeleton (11). One of the heat shock polypeptides of chicken will associate with the src gene product of Rous sarcoma virus (12) and another is found associated with the cytoskeleton (13).

Four of the Drosophila heat shock-activated genes, encoding proteins of M, 27,000 (hsp 27), 26,000 (hsp 26), 23,000 (hsp 23), and 22,000 (hsp 22), are located at chromosomal subdivision 67B. The four genes contained within an 11-kilobase region are not transcribed in the same direction and appear to contain no intervening sequences (14–17). Little information concerning the structure of these four heat shock proteins is available. The size of the proteins has been estimated from NaDodSO₄ gel electrophoresis data, and isoelectric focusing data indicate that hsp 27 and hsp 26 are relatively basic with isoelectric points of \approx 7.5 and hsp 23 and hsp 22 are more acidic with isoelectric points of \approx 5.9 and 5.8, respectively (18, 19). Here we report the primary sequences of the protein coding regions of these four small heat shock genes and the deduced amino acid sequences of the four proteins.

MATERIALS AND METHODS

Restriction enzyme digestions, agarose and acrylamide gel electrophoresis, plasmid DNA isolation, end labeling by T4 polynucleotide kinase, and Maxam and Gilbert DNA sequence analyses were carried out as described (20, 21). Sequence analysis of the *hsp* 27, *hsp* 26, and *hsp* 23 genes was carried out by using DNA from clone J1; analysis of the *hsp* 22 gene was carried out by using DNA from clone T6. J1 and T6 have been described (15).

RESULTS

Primary DNA Sequence of Protein Coding Regions. The primary sequence of the protein coding regions of the hsp 27, hsp 26, hsp 23, and hsp 22 genes has been determined. An overview of the organization of the four heat shock genes at locus 67B is presented in Fig. 1 and the strategy followed in the sequence determination is summarized in Fig. 2. The DNA sequences obtained are shown in Fig. 3. One open reading frame large enough to encode the appropriate heat shock protein was found in each segment of DNA that hybridized to mRNA from heat-treated cells. Previously, we reported the primary sequence of the regions surrounding the 5' end of the mRNAs (20). The ATG that begins each sequence is the first ATG after the proposed site of initiation of transcription identified previously. The open reading frame following the initiation triplet is translated in Fig. 2. All alternative reading frames contain multiple stop codons and could not encode a protein as large as the small heat shock proteins.

Analysis of Predicted Amino Acid Sequences. The predicted amino acid compositions of the four small heat shock proteins are consistent with the limited experimental data available on the proteins. The predicted molecular weights of the proteins based on the amino acid composition (Table 1) match well with the size estimates from NaDodSO₄ gels (17, 18); hsp 27 is 27,091 daltons, hsp 26 is 26,618 daltons, hsp 23 is 23,544 daltons, and hsp 22 is 22,666 daltons.

Radioactive labeling experiments by Arrigo *et al.* (10) demonstrated that hsp 23 was not labeled with $[^{14}C]$ tryptophan whereas the other three small heat shock genes could be labeled with this amino acid. This result indicates that hsp 23 contains no tryptophan and hsp 27, hsp 26, and hsp 22 contain tryptophan. The amino acid composition data in Table 1 are consistent with this finding since only hsp 23 lacks tryptophan.

Comparison of Predicted Amino Acid Sequences. A comparison of the four amino acid sequences revealed that the four are closely related over a large part of their sequences. The four sequences are aligned in Fig. 3; the four amino acid sequences are most homologous from position 85 to position 195. In this region, the same amino acid is used for at least three of the proteins at 71% of the positions, and the same amino acid is used

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

Abbreviation: hsp, heat shock protein.



FIG. 1. General structure of the portion of the 67B cytological locus encoding the four small heat shock proteins. The direction of the arrows indicates the direction of transcription of each mRNA. The length of the arrows corresponds to the size of the transcribed regions. Data are from ref. 15.

for at least two of the proteins at 92% of the positions. A secondary region of homology exists from amino acid positions 1–14, where one amino acid is used by at least three of the proteins at 50% of the positions, and one amino acid is used by at least two of the proteins at 93% of the positions.

Hsp 22 and hsp 23 have deletions of 27 and 20 amino acids relative to hsp 26 and hsp 27; these deletions which occur toward the amino terminus relative to major regions of homologies account for most of the molecular weight differences among the four small heat shock proteins. When aligned as shown in Fig. 2, hsp 22 is missing amino acids 54 through 80, and hsp 23 is missing amino acids 66 through 86. The remainder of the differences in molecular weight are due to truncated carboxy termini. Hsp 22 has 13 fewer amino acids at the carboxy terminus than hsp 27; 9 fewer than hsp 26 and 5 fewer than hsp 23.

The Heat Shock Proteins Are Similar to α -Crystallin. The sequences of the heat shock proteins as presented in Fig. 3 were

compared with known amino acid sequences (24). The amino acid sequence of bovine α -crystallin was found to be similar to that of the heat shock proteins over $\approx 40\%$ of their lengths. The region of greatest homology is shown in Fig. 4 and extends from amino acid 89 to amino acid 164 (Fig. 3) and from amino acid 72 to amino acid 145 (out of 173) in the α -crystallin chain (25, 26). Over this region of 76 amino acids, the same amino acid is used by α -crystallin and at least one of the heat shock genes at 53 (68%) of the positions. Over this same region, the same amino acid is used by α -crystallin and at least three of the heat shock genes at 39 (51%) of the positions. However, in this stretch of 76 amino acids, there are only 59 positions at which the same amino acid is used by at least three of the heat shock genes; α -crystallin matches at least three of the heat shock genes at 39 out of 59 (66%) of these positions. In addition, similar but nonidentical amino acids are found at nine additional positions in this region (111, 116, 128, 140, 143, 148, 152, and 164 in Fig. 4).



FIG. 2. Strategy used to determine the nucleotide sequences of the protein coding regions of the small heat shock genes. Only relevant restriction sites are indicated (for a complete restriction map of the clones, see refs. 22, 20). --, translated region; TERM, termination signal. DNA sequence analysis using the chemical cleavage method of Maxam and Gilbert (23) was carried out from each of the sites indicated (-) at the bottom of the figure. The length of the arrows corresponds to the number of nucleotides actually determined from each start. bp, Base pair(s).

DISCUSSION

The coding regions of the four small heat shock proteins have been identified by primary sequence analysis. Translation from the first ATG of the open reading frame to the first stop codon would produce proteins of M_r 27,100, 26,600, 23,500, and 22,700. These values are in good agreement with the experimentally determined values of 27,000, 26,000, 23,000, and 22,000. Our data predict that only hsp 23 lacks tryptophan, and Arrigo *et al.* (10) found that only hsp 23 could not be labeled with radioactive tryptophan.

The Four Small Heat Shock Proteins Have Similar Sequences. The amino acid sequences of the small heat shock proteins are similar for more than half of their length. The most extensive homology extends from position 85 to position 195

																									- 25
hsp27	1	met ATG	ser TCA	ile ATT	ile ATA	pro CCA	leu CTG	leu CTG	his CAC	leu TTG	ala GCC	arg CGG	glu GAG	leu TTG	asp GAT	his CAT	asp GAC	tyr TAC	arg CGC	thr ACC	asp GAC	trp TGG	glv GGG	his CAT	leu TTG
hsp26	1	met ATG	ser TCG	leu CTA	ser TCT	thr ACT	ser TCG	leu CTT	ser TCG	leu CTT	val GTG	asp GAT	glu GAA	leu CTC		gln CAG	glu GAC	pr o CCC	arg CGC	ser AGC	pr o CCC	ile ATC	tyr TAC	glu CAG	leu CTT
hsp23	met a ATG	ala GCA	asn AAT	ile ATT	pr o CCG	leu TTG	leu TTG	leu TTG	ser AGC	leu CTT	ala GCC	asp GAC	asp Gat	leu TTG	gly GGC	arg CGA	met ATG	ser TCC	met ATG	va 1 CTG	pr o CCC	phe TTC	tyr TAT	glu GAG	leu CCC
hsp22	met a ATG	arg CGT	ser TCC	leu TTA	pr o CCG	met ATG	phe TTT	trp TGG	arg CGC	met ATG	ala GCC	glu GAG	glu GAG	met ATG	ala GCA	arg CGG	met ATG	pro CCA	arg CGC	leu CTC	ser TCC	ser TCG	pr o CCC	phe TTT	his CAC
hsp27	leu CTG	glu GAG	asp GAT	asp GAC	phe TTC	gly GGT	phe TTT	gly GGC	val GTC	his CAT	a la GCC	his CAC	asp Gat	leu CTG	phe TTC	asn AAT	pr o CCG	arg CGT	arg CGC	leu CTG	leu CTA	leu CTG	pro CCC	asu AAC	aa 50 thr ACC
hsp26	gly GGA	leu CTG	gly GGA	leu TTG	his CAT	pr o CCG	his CAT	ser TCC	arg CGC	tyr TAC	va 1 GTG	leu CTG	leu CTG	pro CCC	leu CTT	glv GGC	thr ACT	gln CAG	gln CAG	arg CGC	arg CGT	ser TCC	ile ATC	asn AAC	gly GGA
hsp23	tyr TAC	tyr TAC	cys TGC	g ln CAG	arg CGC	arg CGA	gly CGA	ile ATC	pr o CCC	tyr TAC	leu TTG	ala GCC	leu CTG	val GTT	gly GGA	pro CCG	met ATG	g lu GAG	gln CAG	g ln CAG	leu CTG	arg CGC	gln CAG	leu CTG	glu GAG
hsp22	ala GCC	phe TTC	phe TTC	his CAC	glu GAG	per o CCG	pr o CCC	va 1 GTT	trp TGG	ser AGT	va 1 GTG	ala GCG	leu CTA	pr o CCG	arg AGG	asn AAC		trp TGG	gln CAG	gln CAT	ile ATT	a la GCC	arg CGC	trp TGG	gln CAG
hsp27	leu	gly	leu	gly	arg	arg	arg	tyr	ser	pro	tyr	glu	arg	ser	hi s	gly	pro	hi s	asn	gln	uet	ser	arg	arg	ala
hap26	CTG	GGA	CTG	GGT	CGT	CCT	OGC	TAT	TCG	CCG ser	TAC	GAG	AGG ala	AGC	CAT	GGC	CCC leu	CAC	AAT leu	CAA	ATG	TCA	CGT	CGC	GCG
	TGT	ССТ	TGC	GCA	TCG	ĊĊĠ	ATA	TGC	CCA	TCG	TCG	CCC	GCT	CCC	CAG	GTT	TTG	GCT	TTA	CGG	CGC	GAG	ATG	GCC	AAC
nsp23	198 AAA	g In CAG	GTG	GGC	GCC	ser TCG	TCG	GGA	TCG	ser TCG	GGA	GCC	GTC	TCG	AAA										
hsp22	glu GAG	gln CAG	glu GAG																						
hsp27	ser TCG	gly GGA	g l y GGT	pr o CCA	a sn AAC	ala GCT	leu CTG	leu CTG	pro CCC	ala GCC		va 1 GTG	gly GGC	lys AAA	asp GAT	gly GGC	phe TTC	gln CAG	wal GTG	cys TGC	ile ATC	asp GAT	va 1 GTG	ser TCG	aa 100 gln CAG
hsp27 hsp26	ser TCG arg CGC	gly GGA asn AAC	gly GGT asp GAC	pro CCA ile ATC	asn AAC his CAC	ala GCT trp TGG	leu CTG pro CCG	leu CTG ala GCA	pro CCC thr ACC	ala GCC ala GCC	his CAT	val GTG val GTG	gly GGC glv GGC	lys AAA lys AAG	asp GAT asp GAT	gly GGC gly GGA	phe TTC phe TTC	gln CAG gln CAG	wal GTG wal GTG	cvs TGC cvs TGC	ile ATC met ATG	asp GAT asp GAC	val GTG val GTC	ser TCG ala GCC	aa 100 gln CAG gln CAG
hsp27 hsp26 hsp23	ser TCG arg CGC	gly GGA asn AAC	gly GGT asp GAC	pro CCA ile ATC	asn AAC his CAC	ala GCT trp TGG	leu CTG pro CCG	leu CTG ala GCA	pro CCC thr ACC	ala GCC ala GCC	his CAT	val GTG val GTG ile ATC	gly GGC glv GGC gly GGA	lys AAA lys AAG lys AAG	asp GAT asp GAT asp GAT	gly GGC gly GGA gly GGC	phe TTC phe TTC phe TTC	gln CAG gln CAG gln CAG	val GTG val GTG val GTC	cys TGC cys TGC Cys TGC	ile ATC met ATG met ATG	asp GAT asp GAC asp GAT	val GTG val GTC val GTG	ser TCG ala GCC ser TCG	aa 100 gln CAG gln CAG his CAC
hsp27 hsp26 hsp23 hsp22	ser TCG arg CGC	gly GGA asn AAC	gly GGT asp GAC	pro CCA ile ATC	asn AAC his CAC	ala GCT TGG leu TTG	leu CTG pro CCG pro CCT	leu CTG ala GCA pro CCG	pro CCC thr ACC pro CCG	ala GCC ala GCC ala GCC	his CAT thr ACC	val GTG ile ATC val GTG	glv GGC glv GGC glv GGA asn AAC	lys AAA lys AAG lys AAG lys AAG	asp GAT asp GAT asp GAT asp GAT	glv GGC glv GGA glv GGC glv GGC	phe TTC phe TTC phe TTC tyr TAC	gln CAG gln CAG gln CAG lys AAA	wal GTG wal GTG wal GTC leu CTC	cys TGC cys TGC cys TGC thr ACC	ile ATC met ATG met ATG leu CTG	asp GAT asp GAC asp GAT asp GAC	wal GTC wal GTC wal GTC wal GTC	ser TCG ala GCC ser TCG lys AAG	slo gln CAG gln CAG his CAC asp GAC
hsp27 hsp26 hsp23 hsp22	ser TCG arg CGC	gly GGA asn AAC	gly GGT asp GAC	pro CCA ile ATC	asn AAC his CAC	ala GCT trp TGG leu TTG	leu CTG pro CCG pro CCT	leu CTG ala GCA pro CCG	pro CCC thr ACC pro CCG	ala GCC ala GCC ala GCC	his CAT thr ACC	val GTG Val GTG ile ATC Val GTG	gly GGC glv GGC glv GGA asn AAC	lys AAA lys AAG lys AAG lys AAG	asp GAT asp GAT GAT asp GAT	gly GGC gly GGA gly GGC gly GGC	phe TTC phe TTC phe TTC tyr	gln CAG gln CAG gln CAG lys AAA	wal GTG wal GTC leu CTC	cys TGC cys TGC cys TGC thr ACC	ile ATC met ATG met ATG leu CTG	asp GAT asp GAC asp GAT asp GAC	wal GTG wal GTC wal GTG wal	ser TCG ala GCC ser TCG lys AAG	asp gln CAG gln CAG his CAC asp GAC asp GAC
hap27 hap26 hap23 hap22 hap27	ser TCG arg CGC	gly GGA AAC lys AAG	gly GGT asp GAC Pro CCC	pro CCA ile ATC asn AAC	asn AAC his CAC glu GAG	ala GCT trp TGG leu TTG leu CTG	leu CTG pro CCG pro CCT thr	leu CTG ala GCA pro CCG val GTC	pro CCC thr ACC pro CCG lys AAG	ala GCC ala GCC ala GCC val GTG	his CAT thr ACC wal GTG	val GTG val GTG ile ATC val GTG asp GAC	glv GGC glv GGA asn AAC asn AAC	lys AAA lys AAG lys AAG thr ACC	asp GAT asp GAT asp GAT asp GAT val GTG	glv GGC glv GGA glv GGC glv GGC glv GGC val GGC	phe TTC phe TTC phe TTC tyr	gln CAG gln CAG gln CAG lys AAA val GTA	wal GTG wal GTC wal GTC leu CTC glu GAG	cys TGC cys TGC cys TGC thr ACC gly GGG	ile ATC met ATG met ATG leu CTG lys AAG	asp GAT asp GAC asp GAT asp GAT his CAC	val GTG val GTC val GTC gTC glu GAC	ser TCG ala GCC ser TCG lys AAG glu GAG	asp GAC asp GAC asp GAC acc asp GAC acc acc acc acc acc acc acc acc acc ac
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26	ser TCG arg CGC phe TTC phe TTC	gly GGA asn AAC lys AAG	gly GGT asp GAC pro CCC pro CCC	pro CCA ile ATC asn AAC ser AGT	asn AAC his CAC glu GAG glu GAG	ala GCT trp TGG leu CTG leu CTG leu	leu CTG pro CCG pro CCT thr ACC asn AAC	leu CTG ala GCA pro COG val GTC val	pro CCC thr ACC pro CCG lys AAG	ala GCC ala GCC ala GCC val GTG val GTG	his CAT thr ACC wal GTG	val GTG val GTG ile ATC val GTG GAC asp GAC	gly GGC gly GGA asn AAC ala GCC	lvs AAA lvs AAG lvs AAG lvs AAG thr ACC ser TCC	asp GAT asp GAT asp GAT wal GTG ile ATT	glv GGC glv GGC glv GGC glv GGC glv GGC uel TTG	phe TTC phe TTC tyr TAC	gln CAG gln CAG lys AAA val GTA val GTC	val GTG val GTG val GTC leu CTC glu GAG glu GAG	cys TGC cys TGC cys TGC thr ACC gly GGG gly GGC	ile ATC met ATG leu CTG lys AAG lys	asp GAT asp GAC asp GAT asp GAC his CAC	val GTG val GTC val GTG gTC gLu GAG gLu GAG	ser TCG ala GCC ser TCG lys AAG glu GAG glu GAA	asloo gln CAG his CAC asp GAC asp GAC arg CGC arg CGC
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26 hsp23	ser TCG arg CGC phe TTC phe TTC phe TTC	gly GGA asn AAC lys AAG glu GAG	gly GGT asp GAC pro CCC pro CCC	pro CCA ile ATC asn AAC ser AGT ser AGC	asn AAC his CAC glu GAG glu GAG	ala GCT trp TGG leu TTG leu CTG leu CTG	leu CTG pro CCG pro CCT thr ACC asm AAC val GTG	leu CTG ala GCA pro CCG val GTC val GTC	pro CCC thr ACC pro CCG lys AAG gly GGA	ala GCC ala GCC ala GCC val GTG val GTG	his CAT thr ACC wal GTG gTG gIn CAG	val GTG ile ATC val GTG GTG GAC asp GAC asp GAC	glv GGC glv GGA asn AAC ala GCC asn AAC	lys AAA lys AAG lys AAG lys AAG thr ACC ser TCC	asp GAT asp GAT asp GAT asp GAT torc ile ATT val GTG	glv GGC glv GGA glv GGC glv GGC glv GGC val GTG leu TTG leu GTG	phe TTC phe TTC tyr TAC	gin CAG gin CAG gin CAG ivs AAA vai GTA vai GTC vai	wal GTG wal GTC leu CTC glu GAG glu GAG glu GAG	cys TGC cys TGC cys TGC thr ACC gly GGC gly GGC	ile ATC met ATG leu CTG lys AAG lys AAG asn AAC	asp GAT asp GAC asp GAC his CAC his CAT	val GTG val GTC val GTC val GTC glu GAG glu GAG	ser TCG ala GCC ser TCG lys AAG glu GAA glu GAA glu GAA	as 100 gln CAG his CAC asp GAC arg CGC arg CGC
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26 hsp23 hsp22	phe TTC phe TTC phe TTC tyr TAC	gly GGA asn AAC lys AAG lys AAG glu GAG	gly GGT asp GAC pro CCC pro CCC	pro CCA ile ATC asm AAC ser AGC ser AGC	asn AAC his CAC glu GAG glu GAG glu GAA glu GAG	ala GCT trp TGG leu CTG leu CTG leu CTG leu CTG	leu CTG pro CCG pro CCT thr ACC asn AAC val GTG lys AAG	leu CTG ala GCA pro CCG val GTC val GTC val GTC val	pro CCC thr ACC pro CCG lys AAG gly GGA lys AAG	ala GCC ala GCC ala GCC val GTG val GTG val GTG val	his CAT thr ACC val GTG gIn CAG leu CTG	val GTG ile ATC GTG GTG GAC asp GAC asp GAC asp GAC	glv GGC glv GGC GGA asn AAC ala GCC asn AAC glv GAC	lys AAA lys AAG lys AAG lys AAG thr ACC ser TCC ser TCC ser AGC	asp GAT asp GAT asp GAT asp GAT tasp GAT tasp GAT transformed tasp GAT transformed tasp GAT T S GAT TA T G GAT TA T GAT TA T GAT TA TA T G GAT TA TA TA TA TA TA TA TA TA TA TA TA T	glv GGC glv GGA glv GGC glv GGC glv GGC glv GGC usu TTG leu GTG val gTC	phe TTC phe TTC tyr TAC	gln CAG gln CAG gln CAG lys AAA val GTA val GTC val GTG val	val GTG val GTG val GTC leu CTC glu GAG glu GAG glu GAG	cys TGC cys TGC thr ACC gly GGG gly GGC gly GGC gly GCA	ile ATC met ATG leu CTG lys AAG lys AAG asn AAA	asp GAT asp GAC asp GAC asp GAC his CAC his CAC his CAT his CAT	val GTG val GTC val GTC gTC gLu GAG glu GAG glu GAG	ser TCG ala GCC ser TCG lys AAG glu GAG glu GAA glu GAA glu GAA glu	an 100 gln CAG gln CAG his CAC asp GAC arg CGC arg CGC arg CGC gln CAG
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26 hsp23 hsp22 hsp22	phe TTC phe TTC phe TTC ghe gly GAG	gly GGA asn AAC lys AAG lys AAG glu GAG	gly GGC gCC gro CCC gro CCC gro CCC	pro CCA ile ATC asm AAC ser AGC ser AGC ser AGC	asn AAC his CAC glu GAG glu GAA glu GAA glu GAA	ala GCT trp TGG leu TTG Leu CTG Leu CTG Leu CTG Leu CTG Ret ATG	leu CTG pro CCG pro CCT thr ACC asn AAC yeal ys AAC lys AAC	leu CTG ala GCA pro COG val GTC val GTC val	pro CCC thr ACC pro CCG lys AAG gly GGA lys AAG gly CGA	ala GCC ala GCC ala GCC val GTG val GTG val GTG cTG	his CAT thr ACC wal GTG gIn CAG leu CTG his CAC	val GTG ile ATC wal GTG GAC asp GAC asp GAC asp GAC	gly GGC glv GGA asn AAC ala GCC asn AAC glv GAG val GTG	lys AAA lys AAG lys AAG lys AAG thr ACC ser TCC ser TCC ser AGC	asp GAT asp GAT asp GAT val GTG GTG val GTG val GTT	glv GGC glv GGC glv GGC glv GGC val GTG leu GTG leu TTG tyr TAT	phe TTC phe TTC tyr TAC leu CTG thr ACC	gln CAG gln CAG gln CAG Ivs AAA val GTA val GTG GTG ual CTG	val GTG val GTC leu CTC glu GAG glu GAG glu GAG glu GAG	cys TGC cys TGC thr ACC gly GGC gly GGC gly GGC gly AAG	ile ATC met ATG leu CTG lys AAG asn AAC lys asn AAC lys asn AAA	asp GAT asp GAT asp GAT asp GAT his CAT his CAT his CAT TCG	val GTG GTC val GTC glu GAG glu GAG glu GAG glu GAG glu GAG	ser TCG ala GCC ser TCG alu GAG glu GAG glu GAG glu GAG glu CAG	an 100 gln CAG gln CAG his CAC asp GAC arg CGC arg CGC arg CGC gln CAG arg CGC arg CGC arg CGC arg CGC arg CAG
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26 hsp27 hsp27 hsp27	phe TTC phe TTC phe TTC phe TTC ghe gha gha GAG	gly GGA asn AAC lys AAG glu GAG asp GAC asp	gly GGT asp GAC pro CCC pro CCC gCC gCC gCC gCC gCC gCC gCC	pro CCA ile ATC asn AAC ser AGC ser AGC ser AGC his CAT	asn AAC his CAC glu GAG glu GAG glu GAA glu GAA glu GAA glu GAA glu GAA	ala GCT trp TGG leu CTG leu CTG leu CTG leu CTG leu CTG his CAC	leu CTG pro CCG pro CCT thr ACC asm AAC val GTG lys AAC ile ATC	leu CTG ala GCA pro CCG wal GTC wal GTC wal GTC	pro CCC thr ACC pro CCG lys AAG gly GGA lys AAG gly CAG gln CAG met ATG	ala GCC ala GCC ala GCC val GTG val GTG val GTG val GTG cGT arg GCC	his CAT thr ACC val GTG gIn CAG gln CAG Leu CTG his CAC	val GTG ile ATC val GTG GTG GTG GTG GTG GAC asp GAC asp GAC asp TTT TTT	glv GGC glv GGA asn AAC ala GCC asn AAC glv GAG val GTG val	lys AAA lys AAG lys AAG lys AAG thr ACC ser TCC ser TCC ser CGC arg CGC	asp GAT asp GAT asp GAT asp GAT asp GAT ile ATT val GTG val GTT lys AAG arg CGC	glv GGC glv GGA glv GGC glv GGC glv GGC ual GTG leu GTG tyr TAT tyr	phe TTC phe TTC tyr TAC Leu CTG thr ACC lys AAG	gin CAG gin CAG gin CAG ivs AAA val GTA val GTC val GTG ieu CTG val cTC	vel GTG vel GTG leu CTC glu GAG glu GAG glu GAG glu GAG glu GAG glu CTC	cvs TGC cvs TGC thr ACC glv GGG glv GGC glv GGC glv AAG asp GAT	ile ATC met ATG leu CTG lys AAG lys AAA lys AAA sen AAA glv. GGC glv GGC	asp GAT asp GAC asp GAC his CAC his CAC his CAT tic TCG leu TTG	val GTC val GTC val GTC val GTC glu GAG glu GAG glu GAG glu GAG glu GAG	ala GCC ser TCG ala GCC ser TCG ala GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG coc ala coc coc ala coc coc coc coc coc coc coc coc coc co	as 100 gln CAG gln CAG his CAC asp GAC arg CGC arg CGC arg CGC arg CGC arg CGC arg CGC arg CAG gln CAG arg CAG arg CAG arg CAG asp CAG asp GAC asp GAC asp GAC asp GAC asp GAC asp GAC asp GAC asp GAC asp CAG asp CAC asp GAC asp CAC Asp CAC A Asp CAC A Asp CAC A Asp CAC A Asp CAC A A A A A A A A A A A A A A A A A
hsp27 hsp26 hsp23 hsp22 hsp27 hsp26 hsp27 hsp27 hsp26 hsp23	phe TTC phe TTC phe TTC gly GGG glu GGG glu GGG	gly GGA asn AAC lys AAG glu GAC asp GAC asp GAT	gly GGT asp GAC pro CCC pro CCC gCC gCC gCC gCC gCC gCC gCC gCC gC	pro CCA ile ATC asn AAC ser AGC ser AGC ser AGC ser AGC bis CAT his CAT	asn AAC his CAC glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG	ala GCT trp TGG leu CTG leu CTG leu CTG leu CTG leu CTG his CAC	leu CTG pro CCG pro CCT thr ACC asn AAC thr AAC thr AAC ile ATC ile ATC	leu CTG ala GCA pro CCG val GTC val GTC val GTC	pro CCC thr ACC pro CCG lys AAG glv gGA lys AAG glv CAG gln CAG met ATG thr ACT	ala GCC ala GCC ala GCC val GTG GTG val GTG val GTG cTG arg GCC arg GCC	his CAT thr ACC val GTG gIn CAG gIn CAG leu CTG his CAC his CAC	val GTG ile ATC val GTG asp GAC asp GAC asp GAC asp GAC asp TTT phe TTT phe TTT	gly GGC glv GGA asn AAC ala GCC asn AAC glv GAG yal GTG GTG yal GTG yal	lvs AAA lvs AAG lvs AAG lvs AAG thr ACC ser TCC ser TCC ser CGC arg CGC arg CGC	asp GAT asp GAT asp GAT asp GAT asp GAT ile ATT val GTG val GTG val GTT lvs AAG arg CGC	glv GGC glv GGA glv GGC glv GGC val GGC leu TTG leu GTG val GTG tyr TAT	phe TTC phe TTC tyr TAC Leu CTG thr ACC lys AAG ala GCT	gin CAG gin CAG gin CAG ivs AAA val GTC val GTC val GTC val CTG val cTG cTG	val GTG val GTC leu CTC glu GAG glu GAG glu GAG glu GAG glu CCC pro CCC pro CCA	cvs TGC cvs TGC thr TGC thr GGG glv GGG glv GGC glv GGC glv GGC glv GGC glv CC CC	ile ATC met ATG leu CTG lys AAG lys AAG lys AAG asn AAA glv GGC glv GGC glv GGT	asp GAT asp GAC asp GAC his CAC his CAC his CAT tic TCG leu TTG tyr TAC	val GTG val GTC val GTC glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG glu GAG	ala GCC ser TCG slu GAG glu GAG glu GAA glu GAA glu GAA glu GAA glu GAA ala GCC ala aca caa ac ac	as 100 gln CAG gln CAG his CAC asp GAC arg CGC arg CGC arg CGC gln CAG gln CAG arg CGC gln CAG arg CGC arg CGC gln CAG

																								8	a175
hsp27	lys	WA 1	va 1	ser	thr	val	ser	ser	asp	glv	val	leu	thr	leu	arg	ala	pr o	pe o	pr o	pro	glv	arg	glu	arg	ala
	AAG	GTA	GTG	TCC	ACT	GTC	TCA	TCC	GAG	GCT	GTG	CTG	ACC	CTC	AGG	GCC	CCG	CCG	CCG	CCC	GGC	AGG	GAA	CGG	GCC
hsp26	gln	va 1	va 1	ser	gln	leu	ser	ser	asp	gly	val	leu	thr	va 1	ser	ile	pr o	1ys	pe o	gln	ala	val	glu	asp	lys
	CAA	GTG	GTC	TCG	CAG	CTG	TCG	TCG	GAT	GGC	GTG	CTC	ACC	GTC	AGT	ATT	CCC	AAG	CCG	GAG	GCC	GTC	GAG	GAC	AAG
hsp23	lys	val	ala	ser	thr	leu	ser	ser	asp	gly	ve 1	leu	thr	ile	lys	va 1	pe o	lys	pe o	pr o	ala	ile	glu	asp	lys
	AAG	GTG	GCC	TCC	ACC	TTG	TCC	TCC	GAT	GCT	GTC	CTG	ACC	ATC	AAG	GTG	CCC	AAG	CCA	CCG	GCA	ATC	GAG	Gat	AAG
hsp22	lys	va 1	thr	ser	thr	leu	ser	ser	asp	glv	val	leu	thr	ile	ser	va 1	pr o	asn	pro	PE O	gly	va 1	gln	glu	thr
	AAG	GTG	ACC	TCG	ACG	CTG	AGC	AGC	GAC	GGC	GTT	CTG	ACC	ATC	Agt	GTG	CCC	AAT	CCT	CCA	GGC	GTG	CAG	GAG	ACA
hsp27	arg	ser	glu	arg	ile	val	arg	ile	gln	gln	thr	gly	pro	ala	hls	leu	ser	wa 1	lys	a la	pro	a la	pro	glu	ala
	AGG	TCG	GAG	CGC	ATT	GTC	CGG	ATC	CAG	CAA	ACG	GGG	CCT	GCT	Cat	TTG	AGC	GTC	AAG	GCA	CCG	GCA	CCC	GAG	GCT
hsp26	ser	lys	glu	arg	ile	ile	g ln	ile	g ln	gln	val	glv	pr o	a la	his	leu	a en	wil	lys	a la	ACO	asp	eet	glu	va l
	TCC	AAG	GAG	CGC	ATC	ATT	CAA	ATT	CAG	CAA	GTG	GGA	CCC	GCT	CAC	CTC	AAC	GTT	AAA	GCA	AAT	GAC	AGC	GAG	GTG
hsp23	gly	a en	g lu	arg	ile	wil	g ln	ile	gln	gln	va 1	gl∨	pro	ala	his	leu	a en	wil	lys	g lu	eer	pro	lys	glu	ala
	GGC	AAC	GAG	CGC	ATC	GTT	CAG	ATC	CAG	CAG	GTG	GGA	CCC	GCC	Cat	CTC	AAT	GTG	AAG	GAG	AGT	CCC	AAG	GAG	GCG
hsp22	leu	,1 ys	g lu	arg	glu	wil	thr	ile	glu	gln	thr	glv	glu	per o	ala	lys	lys	ser	ala	glu	glu	PE O	1 78	asp	lys
	CTC	AAG	GAG	CGT	GAG	GTG	ACC	ATC	GAG	CAG	ACT	GGC	GAG	CCG	GCA	AAG	AAG	TCC	GCC	GAG	GAC	CCA	AAA	GAC	AAA
																		_							
hsp27	gly GGC	a sp Gat	gly GGA	lys AAA	ala GCC	g lu GAA	aon AAT	gly GCC	ser AGC	gly GGC	glu GAG	lys AAA	met ATG	g lu GAG	thr ACT	eer Agc	lys AAG	7 TAA							
hsp26	1ys AAG	gly GGC	lys AAG	glu GAG	a en AAC	gly GGA	a la GCA	per o CCCC	a sea AAC	gly GGC	lys AAG	asp GAC	lys AAG	TAA											
hsp23	va 1 GTG	glu GAG	glm CAG	asp GAC	gly GGT	gly GCC	a on AAC	gly GGT	lys AAG	TAG															
hsp22	thr ACC	ala GCC	ser AGT	gln CAG	TAG																				

FIG. 3. Nucleotide sequences of and amino acid sequences encoded by the four small heat shock genes. The ATG at the beginning of each sequence is the first ATG after the proposed site of initiation of transcription identified previously (20). There are multiple stop codons in the other reading frames of each gene. The methionine encoded by *hsp 23* and *hsp 22* was arbitrarily defined as amino acid 1 for the four proteins.

(Fig. 3). The same amino acid is used for at least three of the four proteins at 71% of the positions. A secondary region of homology exists in the first 14 amino acids of the four proteins. The four proteins are relatively dissimilar from amino acid 15 to amino acid 84 and at the carboxyl termini. The former region also includes deletions in hsp 22 and hsp 23 relative to hsp 26

 Table 1. Proposed amino acid compositions of the four small heat shock proteins

Residues per molecule									
Amino acid	hsp 27	hsp 26	hsp 23	hsp 22					
Ala	12	12	11	11					
Arg	19	12	9	10					
Asn	7	8	6	3					
Asp	12	13	11	7					
Cys	1	4	2	0					
Gln	6	11	10	10					
Glu	13	15	16	18					
Gly	22	15	18	10					
His	11	9	5	3					
Ile	5	10	8	3					
Leu	20	18	16	14					
Lys	11	14	10	12					
Met	5	4	5	5					
Phe	6	3	5	6					
Pro	17	15	13	15					
Ser	13	17	14	15					
Thr	11	4	3	9					
Trp	1	1	0	4					
Tyr	4	4	6	4					
Val	17	20	18	15					

and hsp 27. These deletions account for most of the difference in molecular weight between the two groups of proteins.

As expected, the four heat shock genes also have homologous DNA sequences. In the alignment of Fig. 3, from position 85 to position 195, the same nucleotide is found in at least three of the four genes at 77% of the positions and the same nucleotide is found in all four genes at 37% of the positions. The partial homology of these four genes, which are located within an 11kilobase region, suggests that these genes arose from duplication of a single gene.

The similarity among the four small heat shock genes allows the heat shock proteins of *Drosophila* to be grouped into three categories. One group comprises the small heat shock gene family. A second group includes hsp 68 and hsp 70, which have been shown to be partially homologous (27). DNA sequence analysis suggests that hsp 68 and hsp 70 are also similar at the amino acid level (unpublished data). The sole member of the third group is hsp 83. Thus, it is possible that the variety of functions carried out by the products of the heat shock genes is more limited than initially envisioned when it appeared that the synthesis of seven unrelated proteins was induced by heat shock.

 α -Crystallin Is Remarkably Similar to the Small Heat Shock Proteins. A comparison was made between the heat shock proteins and known protein sequences, in hopes that a similarity in sequence might be due to a similarity in function. If a protein of known function were found that was structurally similar to the heat shock proteins, a clue to the function of the heat shock proteins might be provided.

The amino acid sequence of mammalian α -crystallin is surprisingly similar to those of the four small heat shock proteins. For a region containing 76 amino acids, the same amino acid is used by α -crystallin and at least three of the heat shock proteins at 39 positions and at least one of the heat shock proteins

2364	Genetics	: Ingolia and C	raig			Proc. Natl. Acad. Sci. USA 79 (1982)						
	aa 90) 10)0	110	120	130	140	150	160			
α-cry hsp27 hsp26 hsp23 hsp22	S KD KD KD KD	rFsVnLnVKH gFqVcmdVsq gFqVcmdVaq gFqVcmdVsH gFkltLdVKd	FsPeELKVK FkPnELtVK FkPsELnVK FePsELvVg ysELKVK	VLgDVieV) VvdntvvV VvdDsilV VqdnsvlV VLdgsvlV	hGKHEERQDEI eGKHEERgDgi eGKHEERQDdi eGnHEEReDdi gGKsEqqfaE	HGFISREFH HGmIqRhFv HGhImRhFv HGFItRhFv qGgySRhF1	RKYriPadvd RKYtlPkglt RrYlvPdgyk RrYalPpgye RrfvlPegye	PlaiTSsLSS PtkvvStvSS aeqvvSqLSS adkvaStLSS adkvTStLSS	DGVLTV DGVLT1 DGVLTV DGVLT1 DGVLT1			

FIG. 4. Comparison of deduced amino acid sequences of the heat shock proteins and α -crystallin. The sequence of bovine α -crystallin (α -crys) is that of the B₂ chain (25). The numbering is as in Fig. 2 for the heat shock genes. The α -crystallin sequence comprises amino acids 72–145 (out of 173). Matches between α -crystallin and the heat shock proteins are indicated by capital letters. A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

at 53 positions. The homologies are contained in the same 40% of the amino acid sequences, near but not extending to the carboxyl termini. This region of the heat shock proteins that is homologous to α -crystallin is the same region at which the heat shock proteins are most similar to one another. Thus, it seems likely that this region represents a domain that is important for some function of these proteins.

Other intriguing similarities exist between mammalian α crystallins and the four small heat shock genes of *Drosophila* besides the homologies described above. There are four closely related types of α -crystallin $\approx 20,000-30,000$ daltons in size. There are four small heat shock polypeptides, ranging in size from 22,000 to 27,000 daltons. Two of the α -crystallins have isoelectric points of 7.4 and 7.1 and have been designated in the literature as B types (for basic). The other two have isoelectric points of 5.92 and 5.6 and have been designated as A types (for acidic) (28). Two of the heat shock proteins, hsp 27 and hsp 26, are relatively basic and have isoelectric points of \approx 7.5. Hsp 23 and hsp 22 are more acidic and have isoelectric points of \approx 5.9 and 5.8 (18, 19).

 α -Crystallins are a major component of the vertebrate eye lens, comprising 35% of the protein of the lens (for review, see ref. 29). About 30% of the protein being synthesized in the cells of the lens is α -crystallin. α -Crystallins are highly specific polymeric proteins. The four closely related types form aggregates with an average molecular weight of 800,000 (30, 31). These aggregates perform a major structural role in determining the unique properties of the eye lens. A possible reason for the similar amino acid sequence in α -crystallin and the heat shock proteins is that this domain serves to facilitate aggregation. One testable hypothesis is that the heat shock proteins aggregate into complexes similar to the M_r 800,000 α -crystallin complexes. One could imagine that these complexes could serve a structural role in the nucleus, perhaps stabilizing and protecting the DNA against the traumas that are known to induce synthesis of the heat shock proteins.

We thank S. Vernick for expert technical assistance and R. Metzenberg for thoughtful reading of the manuscript. This research was supported by a grant from the U.S. Public Health Service to E.A.C. and by a postdoctoral fellowship from the Leukemia Society of America (T.D.I.). E.A.C. was a recipient of U.S. Public Health Service Research Career Development Award.

- 1. Ashburner, M. & Bonner, J. J. (1979) Cell 17, 244-254.
- Muller, M. J., Xuong, N. & Geiduschek, E. P. (1979) Proc. Natl. Acad. Sci. USA 76, 5222–5225.
- McAlister, L., Strausberg, S., Kulaga, A. & Finkelstein, D. B. (1979) Curr. Genet. 1, 63-74.

- 4. Loomis, W. T. & Wheeler, S. (1980) Dev. Biol. 79, 399-408.
- Barnett, T., Altschuler, M., McDaniel, C. & Mascarenhas, J. (1980) Dev. Genet. 1, 331-340.
- 6. Kelley, P. & Schlessinger, M. (1978) Cell 15, 1277-1286.
- Mitchell, H. K., Moller, G., Peterson, N. S. & Lipps-Sarimento, L. (1979) Dev. Genet. 1, 181-192.
- 8. McAlister, L. & Finkelstein, D. B. (1980) J. Bacteriol. 143, 603-612.
- Velazquez, J. M., DiDomenico, B. J. & Lindquist, S. (1980) Cell 20, 679-689.
- 10. Arrigo, A. P., Fakan, S. & Tissieres, A. (1980) Dev. Biol. 78, 86-103.
- 11. Sinibaldi, R. M. & Morris, P. W. (1981) J. Biol. Chem. 21, 10735-10738.
- Oppermann, H., Levinson, W. & Bishop, J. M. (1981) Proc. Natl. Acad. Sci. USA 78, 1067–1071.
- Wang, C., Gomer, R. H. & Lazarides (1981) Proc. Natl. Acad. Sci. USA 78, 3531–3535.
- Wadsworth, S., Craig, E. A. & McCarthy, B. J. (1980) Proc. Natl. Acad. Sci. USA 77, 2134–2137.
- 15. Craig, E. A. & McCarthy, B. J. (1979) Nucleic Acids Res. 8, 4441-4457.
- Corces, V., Holmgren, R., Freund, R., Morimoto, R. & Meselson, M. (1980) Proc. Natl. Acad. Sci. USA 77, 5390–5393.
- Voellmy, R., Goldschmidt-Clermont, M., Southgate, R., Tissieres, A., Levis, R. & Gehring, W. (1981) Cell 261-271.
 Storti, R. V., Scott, M. P., Rich, A. & Pardue, M. L. (1980) Cell
- Storti, R. V., Scott, M. P., Rich, A. & Pardue, M. L. (1980) Cell
 22, 825–835.
- Mirault, M.-E., Goldschmidt-Clermont, M., Moran, L., Arrigo, P. A. & Tissieres, A. (1978) Cold Spring Harbor Symp. Quant. Biol. 42, 819-827.
- Ingolia, T. D. & Craig, E. A. (1981) Nucleic Acids Res. 9, 1627-1642.
- 21. Craig, E. A., McCarthy, B. J. & Wadsworth, S. C. (1979) Cell 16, 575–585.
- 22. Ingolia, T. D., Craig, E. A. & McCarthy, B. J. (1980) Cell 21, 669-679.
- Maxam; A. & Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 62, 560-564.
- Dayhoff, M. O., Barler, W. C., Hont, L. T. & Schwartz, R. M. (1978) Protein Segment Dictionary 78 (National Biomedical Research Foundation, Washington, D.C.).
- Vander Ouderaa, F. J., de Jong, W. W. & Bloemendal, H. (1973) Eur. J. Biochem. 39, 207-211.
- Vander Ouderaa, F. J., de Jong, W. W. & Bloemendal, H. (1974). Eur. J. Biochem. 49, 157-161.
- Holmgren, R., Livak, K., Morimoto, R., Freund, R. & Meselson, M. (1979) Cell 18, 1359–1370.
- Schoenmakers, J. G. G. & Bloemendal, H. (1968) Nature (London) 220, 790-791.
- 29. Bloemendal, H. (1977) Science 197, 127-138.
- Bloemendal, H., Berns, T., Zweers, A., Hoenders, H. & Bendetti, E. L. (1972) Eur. J. Biochem. 24, 401-409.
- 31. Li, L. K. (1974) Exp. Eye Res. 18, 381-389.