Volume 11 Number 20 1983 Nucleic Acids Research

Localization of the hsp83 transcript within a 3292 nucleotide sequence from the 63B heat shock locus of *D. melanogaster*

Roger W.Hackett and John T.Lis

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853, USA

Received 8 July 1983; Revised and Accepted 15 September 1983

ABSTRACT

We have determined the complete nucleotide sequence of a 3292 bp cloned segment derived from the 63B heat shock cytogenetic locus of D. melanogaster. Within this segment we have positioned the start of transcription and RNA splice sites of the unique gene that encodes the 83,000 d heat shock polypeptide (hsp83 gene) by S1 mapping and synthesis of cDNA from restriction fragment primed mRNA. The sequence begins at a point 879 bp upstream from the transcription start and includes the 149 bp nontranslated first exon, the 1139 bp intron and extends 1125 bp into the protein coding region. These data identify a single open translation reading frame for the first 375 amino acids of the 83,000 d polypeptide, beginning with the first ATG codon located at the 3' intron-exon junction. We discuss and demonstrate the use of E. coli exonuclease III generated single-strand DNA probes as an alternative to strand separation for S1 mapping of mRNA. We also use homology search criteria based upon known protein-DNA binding sites to compare our hsp83 sequence with other sequenced Drosophila heat shock genes. These comparisons indicate that a large region of approximately 80 bp centered around the transcription initiation point of the hsp83 gene shares only a 31% homology with the corresponding region of the hsp70 gene, whereas the hsp22, 23, 26, and 27 genes share a 54% homology with hsp70 in this region. The lower homology of the hsp83 gene is consistent with the deviant nature of this heat shock gene.

INTRODUCTION

Exposure of Drosophila cells to temperatures several degrees above the normal growth temperature of 23°C induces a rapid and dramatic change in gene expression. Most genes active at 23°C are repressed while a specific set of heat shock genes is activated both at the transcriptional and translational levels (1). One of these induced genes, which encodes the 83,000 d heat shock polypeptide (hsp83), is a single copy gene located at the 63B cytogenetic locus of D. melanogaster (2). Several lines of evidence suggest that the expression of the hsp83 gene differs significantly from the other major heat shock genes of D. melanogaster. First, in several continuous cell culture lines the 83,000 d polypeptide (3-5) and RNA (6) are abundant at normal growth temperatures. Second, maximum induction of hsp83 protein synthesis occurs at 33°C, a temperature below optimal for synthesis of other abundant heat shock

proteins (4). Third, in cultured cells grown at 25°C this message is found primarily in poly(A)$^-$ fractions, yet after induction, is found distributed between poly(A)$^+$ and poly(A)$^-$ fractions (3,6). Fourth, larval salivary glands incubated at 25°C produce hsp83 mRNAs that can be isolated and translated in vitro (5), yet that are inefficiently translated in these cells prior to induction (7). Fifth, the 63B region is induced to puff in isolated salivary gland nuclei in response to uninduced cellular extracts that do not cause puffing at other heat shock loci (8). Sixth, hsp83 is a common protein in early embryos kept at normal growth temperatures, and is also the only heat shock protein detectable in heat-induced unfertilized eggs (9). Furthermore, of the seven major heat shock genes only the mRNAs encoding hsp83, hsp27, and hsp26 are present during normal development in non-heat shocked adult ovaries and in preblastoderm embryos. These appear to be maternally derived RNAs synthesized in ovarian nurse cells and transported into oocytes (10). Seventh, hsp83 mRNA is found in uninduced immature ovaries and early egg chambers, whereas other heat shock mRNAs are not detected (10). Finally, the hsp83 gene is the only heat shock gene known to contain an intervening sequence (2). Together, these observations indicate that expression of the hsp83 gene involves regulatory events at the levels of transcription, processing, and translation that are not shared by other heat shock genes.

The basis for the differential expression of the hsp83 gene must ultimately reside in differences in regulatory nucleotide sequences. Therefore, our initial goal was to determine the nucleotide sequence of an interval sufficient to cover potential regulatory regions. Such regions should include the start of transcription and 5' flanking sequences that encompass known sites of DNase I hypersensitivity in chromatin (11) and extend through the start of translation into the protein coding sequence. Identification of these starts requires precise identification of the initiation and splice sites of the hsp83 transcript. Although 578 bp of interrupted DNA sequence from the 5' end of the hsp83 gene and at the splice junctions have been determined (12), the reported sequence omits regions containing the upstream DNaseI hypersensitive sites (11), the hsp83 intron, and fails to establish a translation reading frame for the 83,000 d protein. We present a more extensive, uninterrupted sequence and demonstrate a 375 codon open reading frame. Furthermore, using a rapid exonuclease III/nuclease S1 mRNA mapping procedure, we present high resolution data confirming the previously assigned initiation and 3' splice nucleotides of the hsp83 mRNA. The 5' splice junction is positioned by determining the exact length of cDNA

synthesized from restriction fragment primed heat shock mRNA. We compare sequences flanking the hsp83 gene and the four small heat shock genes at the 67B locus (13) to hsp70 gene flanking sequences (14). The five genes share an extended region of homology, but the homology in the case of the hsp83 gene is less than that shared by the other four genes.

MATERIALS AND METHODS

(a) Enzymes and Substrates

Restriction endonucleases were from Bethesda Research Laboratories or New England Biolabs, *E. coli* exonuclease III and T_4 polynucleotide kinase were from New England Biolabs, nuclease S1 was from Sigma, and AMV reverse transcriptase was a gift from Dr. J. Beard, Life Sciences, Inc. Calf intestinal alkaline phosphatase (Grade I) was obtained from Boehringer Mannheim and purified free of contaminating nucleases as described (15). [α-^{32}P]deoxynucleoside triphosphates and [γ-^{32}P]adenosine triphosphate were from Amersham Corp.

(b) Purification of in vitro labeled DNA fragments

Plasmids were propagated in *E. coli* K12 strains HB101 [F^-, hsdS20 (r_B^-, m_B^-), recA13, ara-14, proA2, lacY1, galK2, rpsL20 (Sm^r), xy-15, mtl-1, supE44, λ^-], (16) or GM33 [F^-, dam-3, sup85], (17), and prepared by equilibrium gradient centrifugation as described (18). DNA preparations were routinely extracted with phenol, RNAse A treated, and precipitated with polyethylene glycol (19). 5'-protruding restriction fragment ends were labeled with T_4 polynucleotide kinase or reverse transcriptase as described (20,21). Flush restriction fragment ends were labeled with high efficiency by either polynucleotide kinase or reverse transcriptase after limited treatment with *E. coli* exonuclease III to generate short (about 5 nucleotide) 5'-protruding ends, as described previously (20). All end-labeled fragments were separated on preparative polyacrylamide gels (20).

(c) DNA sequencing

DNA was sequenced by the Maxam and Gilbert method (21). The sequence for a total of 1859 base pairs, including 1125 base pairs within the protein coding region was determined from both DNA strands. Within this interval we observed 13 positions where band compression (21) required sequence from the opposite strand. From this error frequency of 0.007 we estimate that our sequence for 1433 bases derived from one strand is better than 99% correct.

(d) Isolation of Drosophila RNA

Heat shock poly(A)$^+$ RNA was isolated from polysome pellets of *D.*

melanogaster K_{co} line tissue culture cells, essentially as described (22). Alternatively, poly(A)$^+$ RNA was obtained from D. melanogaster adults after incubating approximately 2000 flies at 37° for 60 minutes, crushing with mortar and pestle in 15 ml ice-cold buffer (10 mM Tris-HCl, pH 8, 20 mM EDTA, 1% SDS, 0.1% diethylpyrocarbonate) and immediately extracting three times with equal volumes of equilibrated phenol and diethyl ether. The RNA was ethanol precipitated and treated with RNase-free DNase I (23) and a poly(A)$^+$ fraction was collected (22).

(e) Exonuclease III/nuclease S1 mRNA mapping

Restriction fragments labeled at one 5' terminus were purified from polyacrylamide gels and digested to completion with an excess of E. coli exonuclease III using the conditions described by Guo and Wu (24). Briefly, purified fragments were treated in 100 µl of exo III buffer (66 mM Tris-HCl, pH 8, 77 mM NaCl, 5 mM $MgCl_2$, 10 mM dithiothreitol) with 1-2 µl of exonuclease III (56 units/µl) at 23°C for 50 minutes. The reaction was terminated by two extractions with an equal volume of equilibrated phenol, followed by ether extraction and ethanol precipitation. Generally about 2 µg of Drosophila heat shock poly(A)$^+$ RNA was added with 23 µg of yeast carrier tRNA to an aliquot of phenol-extracted probe and co-precipitated. Probe aliquots without Drosophila RNA were precipitated with 25 µg of yeast tRNA only. These pellets were then resuspended in 10 µl of S1 hybridization buffer (40 mM PIPES, pH 6.4, 400 mM NaCl, 1 mM EDTA in 20% or 80% formamide), heat-denatured, hybridized, and then treated with nuclease S1 as described (25). The short RNA-DNA hybrid used to position the start of transcription was generated by a hybridization under nonstringent conditions in buffer containing 20% formamide at 30°C for 15 hours. The larger RNA-DNA hybrid used to position the 3' splice junction was generated in 80% formamide at 30°C for 13 hours. Nuclease S1 digestions were in 300 µl at 200 units/ml at 15°C for 60 minutes. Control samples hybridized without Drosophila RNA were incubated in S1 buffer with or without nuclease S1. S1 resistant hybrids were precipitated with isopropanol and loaded onto sequencing gels as described (25).

(f) Restriction fragment primed reverse transcripts

Primer extension was performed by hybridization of primer to RNA at 75°C in S1 hybridization buffer without formamide, and cDNA was synthesized with reverse transcriptase (20). Non-hybridized RNA was digested with RNase A at 250 µg/ml, and after phenol extraction the cDNA was recovered by ethanol precipitation.

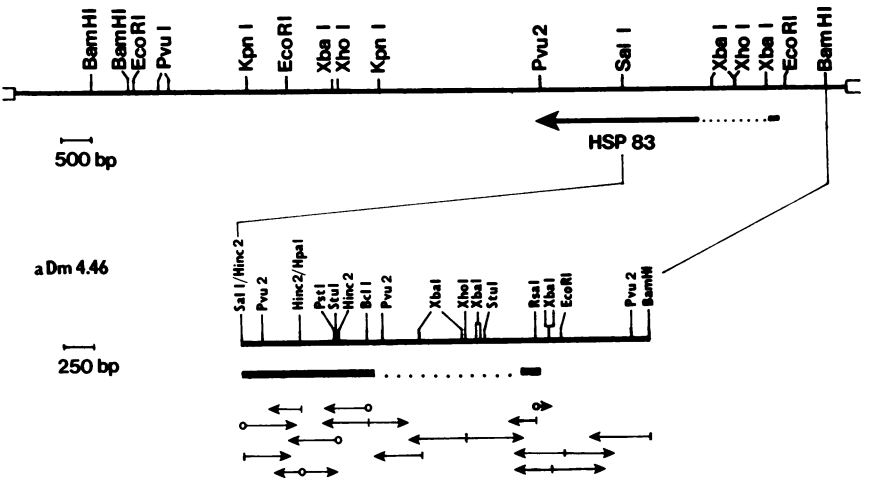


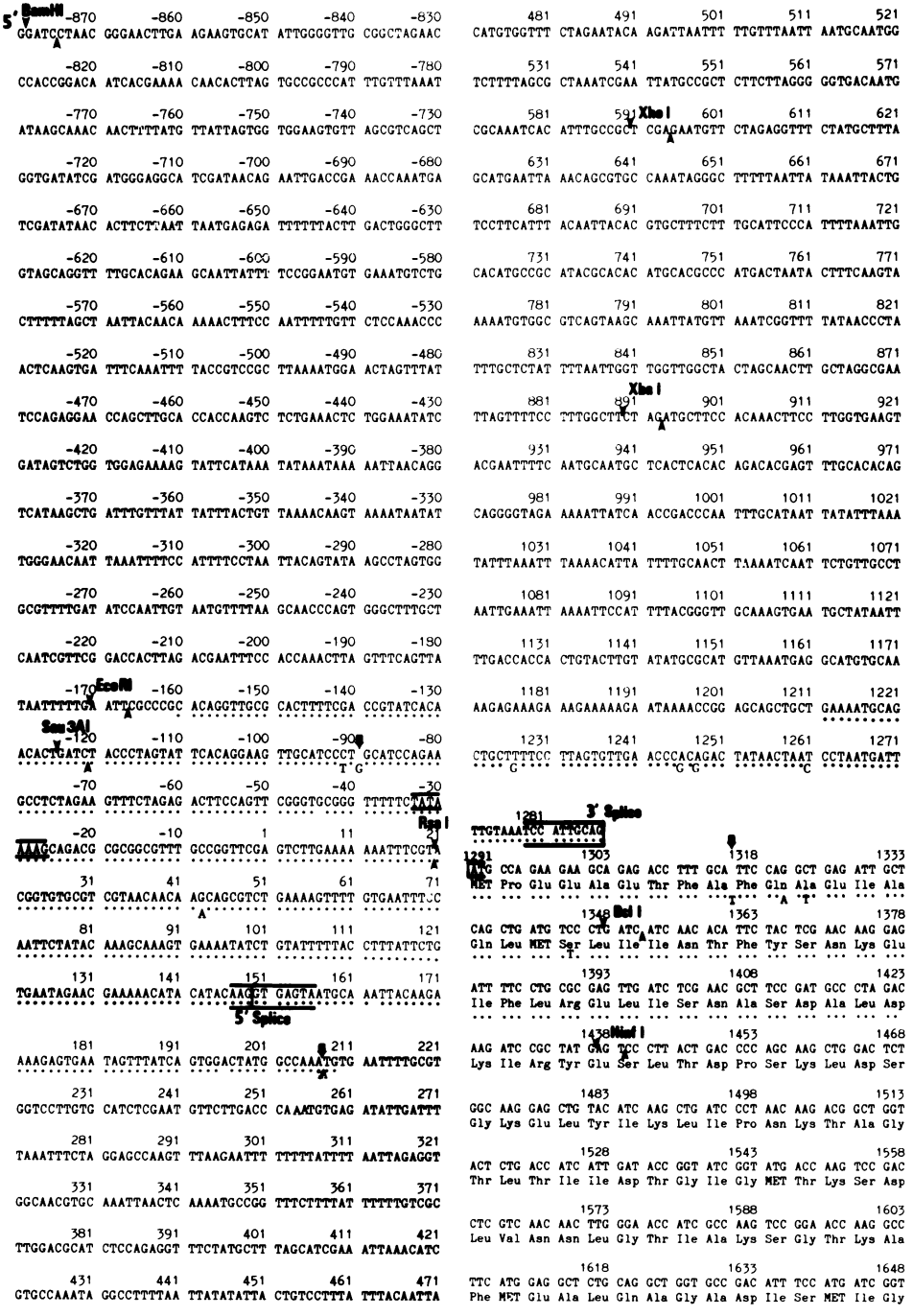
Figure 1. Restriction map of cloned segments from 63B and sequencing strategy.

A low resolution restriction map of the genomic clone λbDm4L (6) carrying 13.3 Kb of D. melanogaster DNA (solid bar) inserted into λ sep6 (open bars) is shown. The solid arrow beneath designates the extent and 5'-3' orientation of the hsp83 transcribed sequences. Intron sequences are indicated by dots. A BamHI-SalI subclone, aDm4.46, is shown in expanded scale with key restriction sites. Each arrow in the sequencing strategy represents the direction and total distance read from a labeled end. Arrows beginning with circles represent 5' kinase end-labeling, and arrows beginning with bars represent 3' labeled ends.

RESULTS

(a) A 3292 bp sequence from the 5' end of the hsp83 gene

Figure 1 illustrates at low resolution the position of the 3292 bp BamHI-SalI insert of subclone aDm4.46 relative to the hsp83 transcript and the original D. melanogaster/λSep6 hybrid phage from which it was isolated. In Figure 2 we present the complete nucleotide sequence derived from this subcloned segment, with key restriction sites, the TATA sequence (26), splice junctions (27), and predicted hsp83 amino acid sequence indicated. Holmgren et al. (12) have reported 578 bp of DNA sequence at the 5' end of a hsp83 gene isolated from a different genomic library, and apparent splice sites of the hsp83 transcript. Figure 2 indicates the regions sequenced by Holmgren et al. (12) and positions where our sequence differs. The nucleotide sequence and splice sites reported previously did not establish an open translation reading frame for the 83,000 d protein, and the first AUG codon within the spliced transcript (12) was followed by translational stop codons. In contrast, our


5' BamHI
-870 -860 -850 -840 -830
GGATCCTAAC GGGAACTTGA AGAAGTGCAT ATTGGGGTTG CGGCTAGAAC
-820 -810 -800 -790 -780
CCACCGGACA ATCACGAAAA CAACACTTAG TGCCGCCCAT TTGTTTAAAT
-770 -760 -750 -740 -730
ATAAGCAAAC AACTTTTATG TTATTAGTGG TGGAAGTGTT AGCGTCAGCT
-720 -710 -700 -690 -680
GGTGATATCG ATGGGAGGCA TCGATAACAG AATTGACCGA AACCAAATGA
-670 -660 -650 -640 -630
TCGATATAAC ACTTCTTAAT TAATGAGAGA TTTTTTACTT GACTGGGCTT
-620 -610 -600 -590 -580
GTAGCAGGTT TTGCACAGAA GCAATTATTT TCCGGAATGT GAAATGTCTG
-570 -560 -550 -540 -530
CTTTTTAGCT AATTACAACA AAAACTTTCC AATTTTGTTT CTCCAAACCC
-520 -510 -500 -490 -480
ACTCAAGTGA TTTCAAATTT TACCGTCCGC TTAAAATGGA ACTAGTTTAT
-470 -460 -450 -440 -430
TCCAGAGGAA CCAGCTTGCA CCACCAAGTC TCTGAAACTC TGGAAATATC
-420 -410 -400 -390 -380
GATAGTCTGG TGGAGAAAAG TATTCATAAA TATAAATAAA AATTAACAGG
-370 -360 -350 -340 -330
TCATAAGCTG ATTTGTTTAT TATTTACTGT TAAAACAAGT AAAATAATAT
-320 -310 -300 -290 -280
TGGGAACAAT TAAATTTTCC ATTTTCCTAA TTACAGTATA AGCCTAGTGG
-270 -260 -250 -240 -230
GCGTTTTGAT ATCCAATTGT AATGTTTTAA GCAACCCAGT GGGCTTTGCT
-220 -210 -200 -190 -180
CAATCGTTCG GACCACTTAG ACGAATTTCC ACCAAACTTA GTTTCAGTTA
-170 EcoRI -160 -150 -140 -130
TAATTTTTGA ATTCGCCCGC ACAGGTTGCG CACTTTTCGA CCGTATCACA
Sau3AI -120 -110 -100 -90 -80
ACACTGATCT ACCCTAGTAT TCACAGGAAG TTGCATCCCT GCATCCAGAA
-70 -60 -50 -40 -30
GCCTCTAGAA GTTTCTAGAG ACTTCCAGTT CGGGTGCGGG TTTTTCTATA
-20 -10 1 11 21
AAAGCAGACG CGCGGCGTTT GCCGGTTCGA GTCTTGAAAA AAATTTCGTA
31 41 51 61 71
CGGTGTGCGT CGTAACAACA AGCAGCGTCT GAAAAGTTTT GTGAATTTCC
81 91 101 111 121
AATTCTATAC AAAGCAAAGT GAAAATATCT GTATTTTTAC CTTTATTCTG
131 141 151 161 171
TGAATAGAAC GAAAAACATA CATACAAGCT GAGTAATGCA AATTACAAGA
5' Splice
181 191 201 211 221
AAAGAGTGAA TAGTTTATCA GTGGACTATG GCCAAATGTG AATTTTGCGT
231 241 251 261 271
GGTCCTTGTG CATCTCGAAT GTTCTTGACC CAAATGTGAG ATATTGATTT
281 291 301 311 321
TAAATTTCTA GGAGCCAAGT TTAAGAATTT TTTTTATTTT AATTAGAGGT
331 341 351 361 371
GGCAACGTGC AAATTAACTC AAAATGCCGG TTTCTTTTAT TTTTTGTCGC
381 391 401 411 421
TTGGACGCAT CTCCAGAGGT TTCTATGCTT TAGCATCGAA ATTAAACATC
431 441 451 461 471
GTGCCAAATA GGCCTTTTAA TTATATATTA CTGTCCTTTA TTTACAATTA
481 491 501 511 521
CATGTGGTTT CTAGAATACA AGATTAATTT TTGTTTAATT AATGCAATGG
531 541 551 561 571
TCTTTTAGCG CTAAATCGAA TTATGCCGCT CTTCTTAGGG GGTGACAATG
581 591 XhoI 601 611 621
CGCAAATCAC ATTTGCCGCT CGAGAATGTT CTAGAGGTTT CTATGCTTTA
631 641 651 661 671
GCATGAATTA AACAGCGTGC CAAATAGGGC TTTTTAATTA TAAATTACTG
681 691 701 711 721
TCCTTCATTT ACAATTACAC GTGCTTTCTT TGCATTCCCA TTTTAAATTG
731 741 751 761 771
CACATGCCGC ATACGCACAC ATGCACGCCC ATGACTAATA CTTTCAAGTA
781 791 801 811 821
AAAATGTGGC GTCAGTAAGC AAATTATGTT AAATCGGTTT TATAACCCTA
831 841 851 861 871
TTTGCTCTAT TTTAATTGGT TGGTTGGCTA CTAGCAACTT GCTAGGCGAA
XhoI
881 891 901 911 921
TTAGTTTTCC TTTGGCTTCT AGATGCTTCC ACAAACTTCC TTGGTGAAGT
931 941 951 961 971
ACGAATTTTC AATGCAATGC TCACTCACAC AGACACGAGT TTGCACACAG
981 991 1001 1011 1021
CAGGGGTAGA AAAATTATCA ACCGACCCAA TTTGCATAAT TATATTTAAA
1031 1041 1051 1061 1071
TATTTAAATT TAAAACATTA TTTTGCAACT TAAAATCAAT TCTGTTGCCT
1081 1091 1101 1111 1121
AATTGAAATT AAAATTCCAT TTTACGGGTT GCAAAGTGAA TGCTATAATT
1131 1141 1151 1161 1171
TTGACCACCA CTGTACTTGT ATATGCGCAT GTTAAATGAG GCATGTGCAA
1181 1191 1201 1211 1221
AAGAGAAAGA AAGAAAAAGA ATAAAACCGG AGCAGCTGCT GAAAATGCAG
1231 1241 1251 1261 1271
CTGCTTTTCC TTAGTGTTGA ACCCACAGAC TATAACTAAT CCTAATGATT
3' Splice
1281
TTGTAAATCC ATTGCAG
1291 1303 1318 1333
TTG CCA GAA GAA GCA GAG ACC TTT GCA TTC CAG GCT GAG ATT GCT
MET Pro Glu Glu Ala Glu Thr Phe Ala Phe Gln Ala Glu Ile Ala
1348 BclI 1363 1378
CAG CTG ATG TCC CTG ATC ATC AAC ACA TTC TAC TCG AAC AAG GAG
Gln Leu MET Ser Leu Ile Ile Asn Thr Phe Tyr Ser Asn Lys Glu
1393 1408 1423
ATT TTC CTG CGC GAG TTG ATC TCG AAC GCT TCC GAT GCC CTA GAC
Ile Phe Leu Arg Glu Leu Ile Ser Asn Ala Ser Asp Ala Leu Asp
1438 NlaIII 1453 1468
AAG ATC CGC TAT GAG TCC CTT ACT GAC CCC AGC AAG CTG GAC TCT
Lys Ile Arg Tyr Glu Ser Leu Thr Asp Pro Ser Lys Leu Asp Ser
1483 1498 1513
GGC AAG GAG CTG TAC ATC AAG CTG ATC CCT AAC AAG ACG GCT GGT
Gly Lys Glu Leu Tyr Ile Lys Leu Ile Pro Asn Lys Thr Ala Gly
1528 1543 1558
ACT CTG ACC ATC ATT GAT ACC GGT ATC GGT ATG ACC AAG TCC GAC
Thr Leu Thr Ile Ile Asp Thr Gly Ile Gly MET Thr Lys Ser Asp
1573 1588 1603
CTC GTC AAC AAC TTG GGA ACC ATC GCC AAG TCC GGA ACC AAG GCC
Leu Val Asn Asn Leu Gly Thr Ile Ala Lys Ser Gly Thr Lys Ala
1618 1633 1648
TTC ATG GAG GCT CTG CAG GCT GGT GCC GAC ATT TCC ATG ATC GGT
Phe MET Glu Ala Leu Gln Ala Gly Ala Asp Ile Ser MET Ile Gly

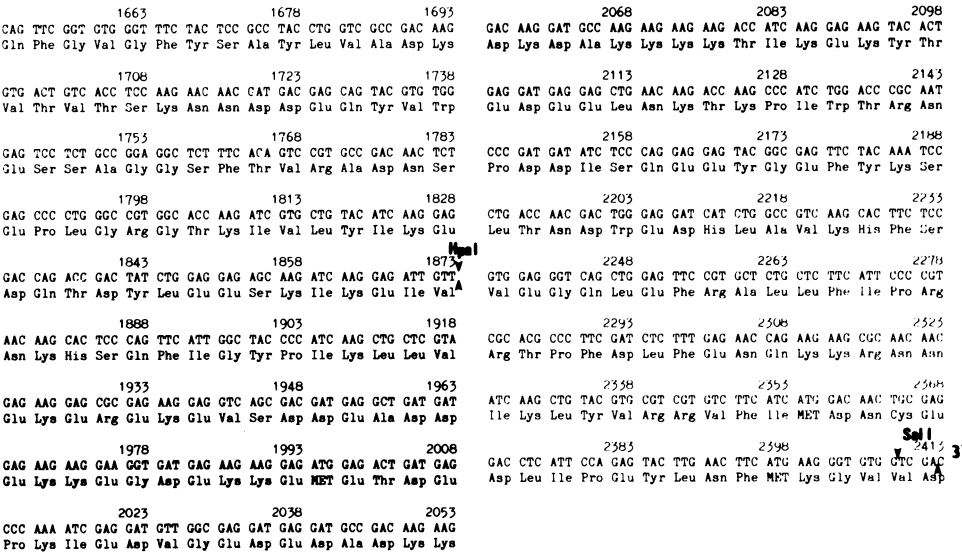


Figure 2. Nucleotide sequence of the 3292 bp aDm4.46 insert.
The nucleotide sequence is shown oriented 5' to 3' from the BamHI site. Nucleotide +1 designates the start of hsp83 transcription. Several restriction sites from Figure 1 and the text are included. Solid boxes highlight the TATA sequence, the 5' splice junction, and the 3' splice junction. The unique open translation reading frame begins immediately adjacent to the 3' splice junction, and the predicted amino acids are shown. Positions where our sequence overlaps and agrees with that of Holmgren et al. (1981) are indicated by dots. At positions where the sequences differ, we show their nucleotides below ours. Arrows indicate bases reported by Holmgren et al. (1981) not found in our sequence.

DNA sequence contains two in-phase translational start codons, at positions +1289 and +1340 near the 3' splice junction, that could serve as the initiating AUG and specify the same open reading frame. In order to assign the start of translation to one of these, a precise determination of the 3' splice junction was required. Below we describe transcript mapping experiments that provide high resolution confirmation of the previously predicted initiation and splice sites of the hsp83 transcript, and which taken together with our DNA sequence, establish an open translation reading frame beginning immediately adjacent to the splice site.

(b) Determination of the 3' splice junction

To generate 5' terminally-labeled, single-strand probes for nuclease S1 transcript mapping, we have used the strategy shown in Figure 3(a). This

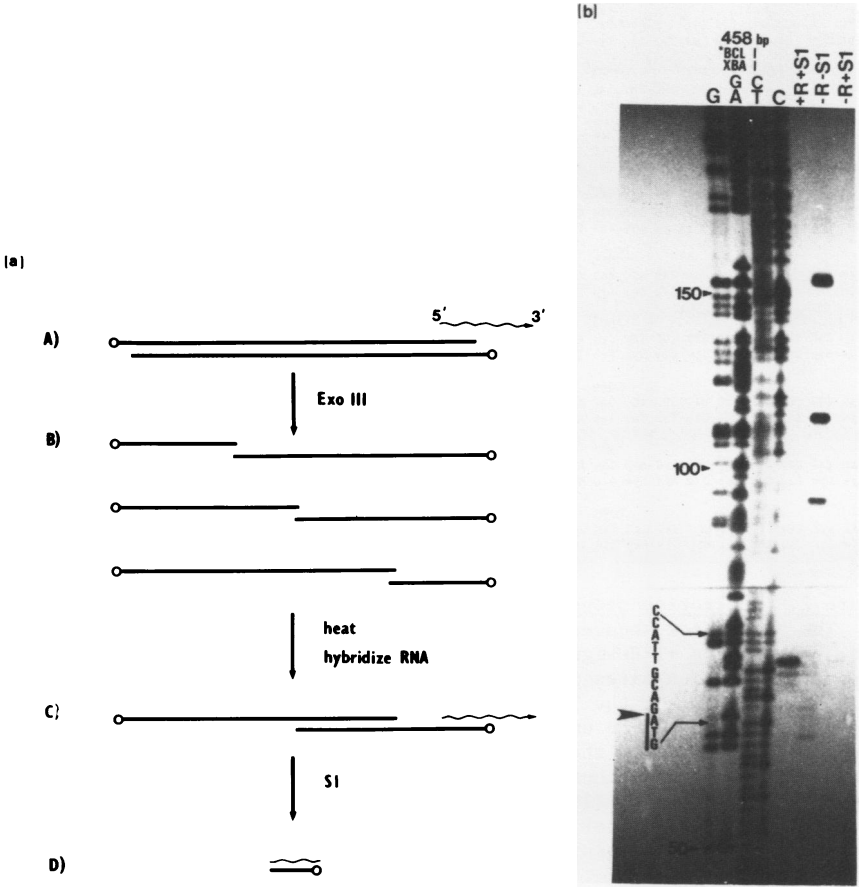


Figure 3.

(a) Nuclease S1 transcript mapping using exonuclease III generated single-strand probes.

A. A restriction fragment (solid lines) with 5' protruding ends (blunt end fragments also may be used) is shown with the transcript to be mapped (wavy lines). Open circles represent 5' phosphates. Both 5' termini may be labeled or just the one needed for mapping.

B. Digestion with E. coli exonuclease III yields molecules that have been more or less synchronously digested from both 3' ends.

C. The mixture of single-strand DNAs is heat denatured and hybridized to RNA. Competition from opposite DNA strands is eliminated. Any DNA reannealing does not interfere with hybridization to RNA, and results in unlabeled DNA fragments after S1 treatment.

D. Digestion with nuclease S1 leaves the RNA-DNA hybrid as the only labeled fragment.

(b) Nuclease S1 mapping of the 3' splice junction.

A 458 bp BclI-XbaI restriction fragment from aDm4.46, kinase end-labeled at the BclI site, was treated with exonuclease III and hybridized to heat shock, poly(A)$^+$ RNA. This same fragment was sequenced to provide exact size standards. Numbers designate distance

from the terminal labeled nucleotide. Hybridizations were with (+R) or without (-R) Drosophila RNA, and then treated with (+S1) or without (-S1) nuclease S1. A short stretch of sequence deduced for the opposite strand is aligned with the sequencing ladder at the splice point. The actual splice site, in agreement with consensus sequences (27), is indicated by an arrow. The first ATG codon is overlined. Samples were run on an 80 cm, 6% sequencing gel.

approach exploits the duplex-specific and processive digestion by E. coli exonuclease III of 3' hydroxy termini (24). Under controlled digestion conditions at 23°C, exonuclease III synchronously removes approximately 10 nucleotides per minute from each 3' hydroxy terminus and exhibits little sequence dependence (24). This exonuclease III digest provides a means of generating in high yield a single-stranded probe that is more rapid than strand separation.

As shown in Figure 3(b), exonuclease III digestion of the BclI-XbaI fragment (+890 to +1348 in Figure 2), which is labeled at the BclI site, produces an array of single-stranded fragments (-RNA -S1 lane) that yield no S1-resistant duplexes after hybridization in the absence of Drosophila RNA (-RNA +S1 lane). Addition of heat shock poly(A)$^+$ Drosophila RNA to the hybridization mixture results in two S1-resistant fragments (+RNA +S1 lane). Since S1 generated, 5' end-labeled fragments migrate 1.5 nucleotides slower on high resolution sequencing gels than comparable Maxam-Gilbert cleavage fragments (28), these bands represent S1-resistant fragments that are 67 and 65 nucleotides long. This places the 3' splice junction at positions +1285 to +1287 in Figure 2. These S1 termini occur within a sequence (5'-TCCATTGCAGAT-3') that perfectly matches the consensus Drosophila intron-exon boundary sequence (5'-TCPy$^{A}_{N}$PyPy$^{G\ C}_{PyT}$AG/$^{Pu}_{C}$T-3') (27) and the more general eucaryotic consensus (29). Holmgren et al. (12), using different conditions of nuclease S1 digestion, also localized the 3' splice junction to this consensus sequence. Since nuclease S1 may leave short overhangs (30), we position the 3' splice site at position +1288 in accordance with the established consensus splice sequence. This location agrees well with previous electron microscopic measurements of hsp83 R-loops (6). The adjacent ATG codon at +1289 is the first initiation codon of a long open reading frame, as we shall demonstrate.

(c) Determination of the hsp83 mRNA start

Figure 4 shows that the complete exonuclease III digestion of an end-labeled RsaI-Sau3AI fragment (-116 to +24) generates a series of single-strand probes (-RNA -S1 lane) that do not give rise to any S1-resistant

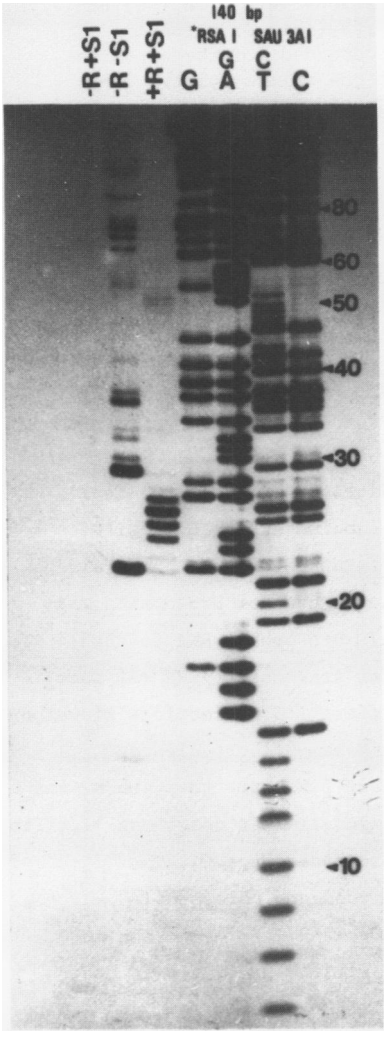


Figure 4. Nuclease S1 mapping of the hsp83 mRNA start. A 140 bp RsaI-Sau3AI restriction fragment from aDm4.46, kinase end-labeled at the RsaI site, was digested with exonuclease III and hybridized to heat shock, poly(A)$^+$ RNA. This same fragment was sequenced to provide exact size standards. Numbers along the right margin indicate distance from the labeled nucleotide. Symbols for each lane are as in Figure 3(b). Samples were run on a 40 cm, 15% sequencing gel.

duplexes when hybridized without Drosophila RNA (-RNA +S1 lane). Under the same hybridization conditions with Drosophila heat shock poly(A)$^+$ RNA, specific S1-resistant fragments are detected (+RNA +S1 lane). Again, 5' end-labeled S1 fragments migrate 1.5 nucleotides slower than their chemical cleavage counterparts (28). Therefore, the protected fragments vary from 20 nucleotides to 25 nucleotides in length. The 20 nucleotide S1-resistant fragment corresponds to an adenosine residue located within a sequence (5'-CGAGTCT-3') that partly matches the eucaryotic consensus mRNA capping site

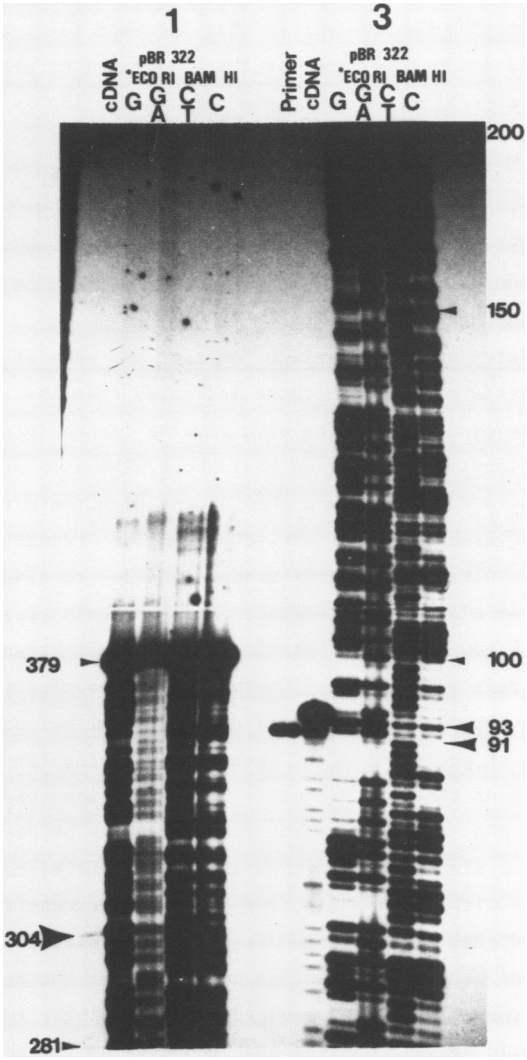


Figure 5. Determination of the 5' splice junction.
The staggered first and third loadings on an 80 cm, 6% sequencing gel are shown. The lane labeled "Primer" contains the 90 nucleotide 3' end-labeled strand of the BclI-HinfI fragment from aDm4.46. The cDNA species primed from heat shock Drosophila RNA (cDNA lanes) is seen in the first loading. This same sample in the third loading shows the original primer strand, as well as a species two nucleotides larger generated by reannealing of the primer restriction fragment and reverse transcriptase repair of the BclI end. The sequencing standard is the 375 bp EcoRI-BamHI fragment from pBR322, 3' end-labeled at the the EcoRI site with [α-^{32}P]dATP. Numbers indicate distance from the terminal labeled nucleotide at the EcoRI site.

(5'-PyCATTCPu-3') determined previously (31). Because many eucaryotic mRNAs initiate at an adenosine residue within this sequence (32), and because comparably situated adenosines occur at the 5' termini of other heat shock genes in Drosophila (12,13), we designate this nucleotide as +1 in Figure 2. Holmgren et al. (12) also detect an S1-resistant terminus corresponding to position -4 in our sequence; the additional fragments seen in Figure 4 probably represent short 3' DNA overhangs not removed by S1 nuclease (30). The hsp83 mRNA start at nucleotide +1 is also located an appropriate distance, 26 nucleotides, from a sequence (5'-TATAAAA-3') that matches the consensus eucaryotic TATA sequence (26).

(d) Determination of the 5' splice junction and size of the small exon

Previous analysis of the hsp83 gene indicated that the single intron is preceded by a short exon (2,12). We used reverse transcriptase to generate cDNA from restriction fragment primed heat shock mRNA as a means to determine the size of the small exon. Plasmid aDm4.46 was digested with BclI and 3' end-labeled with reverse transcriptase, dGTP and [α-^{32}P]dATP, and then cleaved with HinfI. The resulting BclI-HinfI fragment (+1347 to +1436) was gel purified and used as a primer to initiate cDNA synthesis after hybridization to hsp83 mRNA. The length of the RNA homologous primer strand from the HinfI terminus to the partially filled BclI terminus is 90 nucleotides. Reverse transcripts primed from BclI will extend an additional 61 nucleotides to the 3' splice junction at position +1288, thereby accounting for 151 nucleotides of the final length of the cDNA product. By subtracting this length from the total size of the cDNA, the size of the small exon may be determined. As shown in Figure 5, a single cDNA species is evident and migrates on an 80 cm long sequencing gel with a 304 nucleotide chemical cleavage fragment derived from an end-labeled pBR322 restriction fragment. We used the 375 bp EcoRI-BamHI fragment of plasmid pBR322, end-labeled at the EcoRI site with [α-^{32}P]dATP, as a convenient size standard since no single restriction fragment from plasmid aDm4.46 has the same sequence as the cDNA species. However, we have also compared the mobility of this cDNA on 80 cm gels relative to a sequenced, 3' end-labeled HpaI-SalI fragment from aDm4.46, and find an apparent mobility of 302 nucleotides (not shown). This difference may be explained by the slower mobility on high resolution sequencing gels of single strands with higher T content (33). Thus, the large pBR322 chemical cleavage fragments, with 19% T content, move faster relative to the Drosophila chemical cleavage fragments containing 29% T. Since the cDNA sequence has a 32% T content, we use 302 nucleotides as the apparent total length of the

cDNA. In addition, because chemical cleavage products of 3' end-labeled DNA migrate 1.0 nucleotide faster than reverse transcriptase generated fragments (28), the actual length of the cDNA species is 301 nucleotides, and thus the small hsp83 exon is 150 nucleotides long (301-151 nucleotides). From position +1 in our sequence, this places the 5' splice junction at position +150 in Figure 2. This site occurs within a sequence (5'-AAGGTGAGTA-3') that perfectly matches both the eucaryotic (5'-$^{C}_{A}$AG/GT$^{A}_{G}$AGT-3') exon-intron consensus boundary sequence (29), and a consensus sequence for Drosophila (5'-$^{G}_{Pu}$/GTPu$^{A}_{C}$GTPu-3') exon-intron junctions (27). From this match to consensus splice sequences we conclude that the small exon of the hsp83 gene is 149 nucleotides in length. This value is in accord with previous determinations that estimated this exon to be 0.16 Kb long, based upon DNA-primed reverse transcripts, and closer to 0.15 Kb, based upon consensus sequence matches in this region of the hsp83 gene (12).

(e) Identification and evaluation of a translation reading frame for hsp83

Examination of sequences contained within the spliced hsp83 transcript, derived from the S1 nuclease and cDNA results, indicates that the first ATG codon downstream from the mRNA initiation nucleotide occurs at position +1289, immediately after the 3' intron-exon junction. This ATG codon specifies an open translation reading frame that extends through the remaining 1122 nucleotides in the DNA sequence, and is capable of encoding the first 375 amino acids of the 83,000 d polypeptide. The sequence at the first ATG codon (5'-CAAG/ATGC-3') matches in part the eucaryotic translation start consensus (34). Moreover, this ATG specifies the only open reading frame in our sequence. Although a second ATG codon in this same reading frame occurs at position +1340 in our DNA sequence, we adopt the first ATG codon in the 5'-noncoding region as the start of translation. In Figure 2 the first 375 amino acids predicted for the amino half of the hsp83 protein are shown. These amino acids account for about 41,250 d of the total 83,000 d determined by SDS-polyacrylamide gel electrophoresis (2). From our DNA sequence we note that the amino terminal half of hsp83, particularly between amino acids 210 and 275, should be rich in acidic and basic (hydrophilic) residues. Indeed, of the 64 amino acids specified in this region, 48 are glutamic acid, aspartic acid, lysine or arginine residues arranged as strings of hydrophilic groups. Such a composition of charged amino acids is consistent with the known cytoplasmic location of hsp83 (35). Finally, we have compared our hsp83 amino acid sequence to the current protein sequence database, as described by Barker

and Dayhoff (36). No significant homology was found. However, comparison to the yeast hsp90 gene nucleotide sequence determined by F. Farrelly and D. Finkelstein (personal communication) reveals four long (38-103 amino acid) stretches that are 60-90% conserved. The first stretch begins at codon 6 in the Drosophila hsp83 open reading frame.

(f) Comparison of the hsp83 nucleotide sequence to other Drosophila genes--a new twist in the method of searching for common sequences among different genes

We have used a package of computer programs (37) designed to identify homologies between DNA sequences and display the results in dot-matrix form. When a homology of the specified length and percent match is identified, a dot is printed on the graph whose (x,y) coordinates correspond to the position of the first nucleotide of the homology within each sequence.

The criteria used to search for homologies were derived from known procaryotic protein-DNA binding sites. We have examined two instances in E. coli where DNA binding sites for the same regulatory protein occur at several genes or operons that display similar regulation. Five different sites for lexA repressor binding have an average 64% match over twenty nucleotides (38). Five cAMP-CRP binding sites exhibit an average match of 50% over 24 nucleotides (39). In our comparisons of Drosophila flanking sequences, using the criteria of a minimum 60% match between 20 nucleotide blocks yields plots in which known homologies such as TATA sequences are detected, while random matches are not excessive.

DNA sequence data that include 5' flanking regions are available for the hsp70 (14,40), hsp68 (12), and hsp22, hsp23, hsp26, and hsp27 (13,41) genes of D. melanogaster. Because the hsp70 and hsp68 genes are very similar and probably reflect a relatively recent duplication event (2), we have omitted the hsp68 gene in the figures, though it shows striking homology to hsp70 in the region from -42 to +30 (12). Of the five copies of hsp70 (42), we use the sequence of the distal copy at the 87A7 locus (14) as the standard. A deletion ending upstream of this copy demonstrates that sequences necessary for heat induced expression in vivo do not extend beyond 479 nucleotides upstream from the start of transcription (43).

We have superimposed dot-matrix plots in order to illustrate homology between each of the heat shock genes and hsp70. We find three striking features in this matrix. First, the most conspicuous is an extended diagonal line representing a nested set of homologies at nearly identical positions surrounding the mRNA starts shown on an expanded scale in Figure 6. Second, a

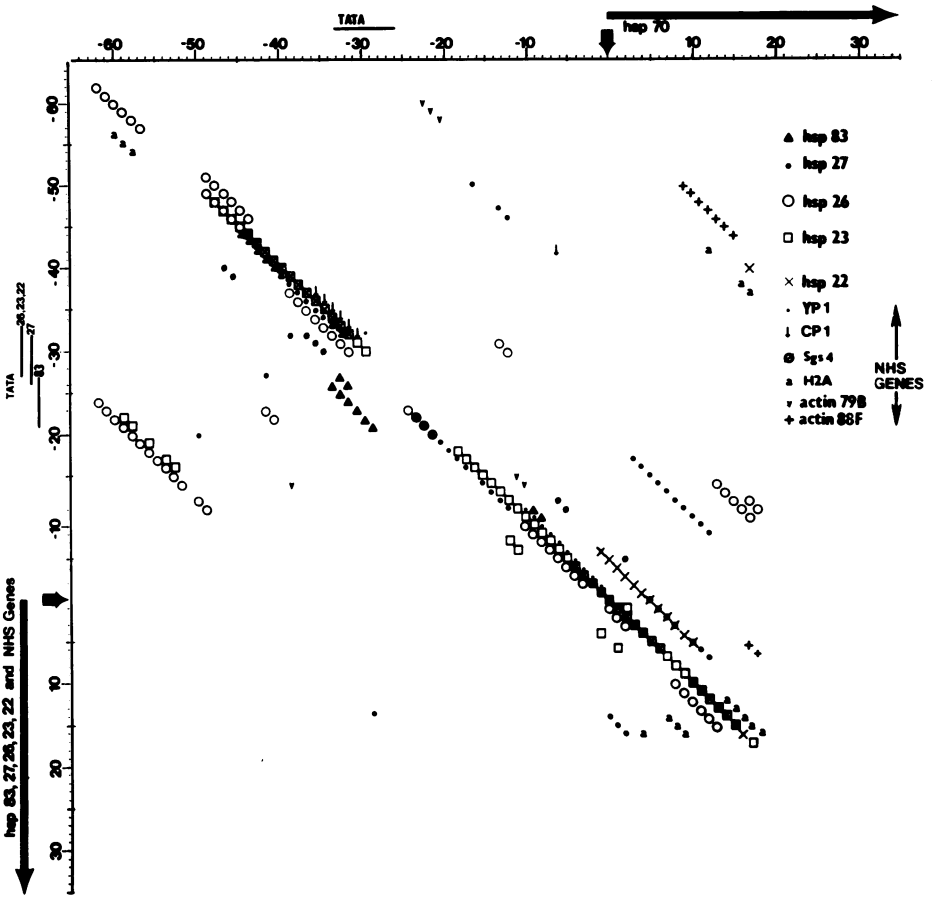


Figure 6.

Expanded GRAPHOM outputs for five heat shock gene sequences compared to hsp70 sequences.

All five genes have been aligned at their transcription starts (+1). Criteria used to generate these plots were 20 nucleotide blocks with at least 60% match. Only 5' flanking and noncoding sequences have been compared, and the hsp83 intron has been deleted. Arrows indicate the transcribed sequences. We show similar comparisons to hsp70, using the same criteria, of sequences flanking other Drosophila genes that are not heat shock inducible (NHS genes), including yolk protein (52), cuticle protein (27), salivary glue protein (53), histone (54), and two actin (55) genes.

block of clustered homologies centered around coordinates (-150, -100, not included in Figure 6) is almost entirely due to repeated sequences common to hsp70 and hsp26 flanking regions. Third, a larger block of scattered

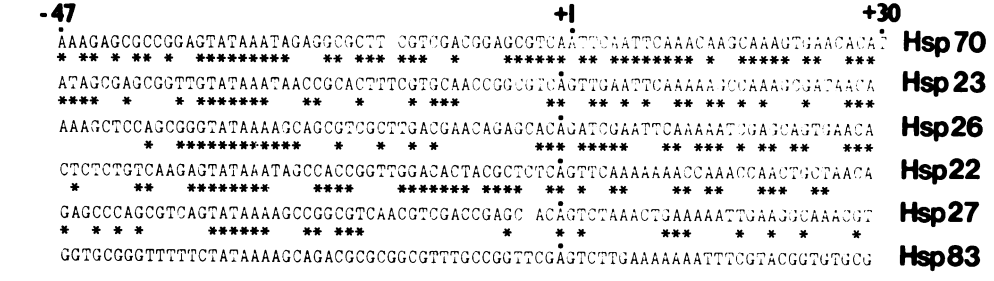


Figure 7. Alignment of heat shock gene homologous flanking sequences.
The standard hsp70 sequence (14) from the distal copy at 87A7 is shown at top. All other sequences have been aligned with this sequence at the TATA homology. Numbering is from the start of transcription at +1. We have inserted a single base loopout in the hsp70 and hsp27 sequences, indicated by gaps, to permit a nonoverlapping display of the homologies from Figure 6. The matches to hsp70 are indicated by asterisks above each sequence. For hsp83, the six overlapping matches shown in Figure 6 do not align the TATA consensus sequences and are omitted here. Small heat shock gene sequences are from Ingolia and Craig (13).

homologies occurs between the transcribed, 5' noncoding sequences of all five genes and hsp70 (centered at +150 on hsp70). Interestingly, the hsp83 leader RNA sequence contributes fewer homologies than other heat shock genes.

The homologies arising from a particular gene are identified by symbols in Figure 6. The small heat shock gene sequences shares significant homology with the hsp70 gene. By comparison, the hsp83 and hsp70 genes share homology only at the TATA sequence. We also note similarities to other Drosophila genes that encode abundant non-heat shock RNAs, though the matches never equal those between the small heat shock genes and the hsp70 gene.

The nucleotide sequence of the hsp70 gene and the corresponding matches in the other heat shock genes are shown in Figure 7. The matches identified by the matrix analysis are indicated by asterisks above each sequence. Introducing a single nucleotide loopout in the hsp70 and hsp27 sequences allows us to present these matches in linear, nonoverlapping form. The two loopouts serve only to better align regions of homology already noted in the dot-matrix and do not alter the overall pattern of homologies. Examination of the aligned sequences shows that the percent match with hsp70 in this 77 bp interval is 47% for hsp26, 49% for hsp22, 52% for hsp27, and 68% for hsp23.

The small heat shock genes and the hsp70 gene in D. melanogaster have conserved flanking sequences that lie roughly between nucleotides -47 and +30 in each gene. Furthermore, the average match to hsp70 sequences within these

regions is 54%. These larger regions of lower average homology to hsp70 gene flanking sequences have not been noted in previous sequence comparisons (12,13) that focused on short (3-12 nucleotide), highly conserved sequence blocks. Surprisingly, in similar comparisons we find that the small heat shock genes, which are related by a gene duplication event, show more homology to the hsp70 gene in this region than to each other. In contrast to the small heat shock genes, sequences flanking the hsp83 gene do not exhibit extensive homology to hsp70 sequences. The match to hsp70 sequences within this 77 nucleotide region is only 31%. The TATA sequence forms the only homology longer than three continuous nucleotides. We conclude that sequences flanking the hsp83 gene are homologous only to a subset of those hsp70 sequences shared by the small heat shock genes.

DISCUSSION

We have sequenced a 3292 bp region that spans the 5' DNAse I hypersensitive sites, the transcription start, the entire first exon, the intron, and translation start of the hsp83 gene. The transcription start and splice sites were mapped at high resolution and the results confirm previously made predictions (2,12). We assign the translation start to the first nucleotide of exon 2, as predicted from the open reading frame that begins at this ATG and extends at least 1125 bp. With these functional landmarks defined we have compared the hsp83 sequence to other heat shock gene sequences. Like the hsp70 and small heat shock genes, the 5' noncoding portion of mature hsp83 mRNA is long, 149 nucleotides. Unlike the other heat shock genes, this region is interrupted by an intron, possesses a lower adenosine content which is typical of some nonheat shock transcripts (13), and has the least homology with hsp70 of any major heat shock gene.

In an attempt to identify potential regulatory sequences in regions flanking several heat shock genes, we have conducted homology searches using criteria based upon defined procaryotic DNA-protein interactions. We find in all cases except hsp83 that sequences surrounding the transcription starts exhibit homology. How well does this flanking sequence homology agree with deletion analysis of the hsp70 gene? Genetic evidence from 87A deficiencies limits the maximum amount of 5' flanking sequence required for regulated expression *in vivo* to 479 nucleotides (43). *In vitro* generated deletions assayed in heterologous systems suggest that only about 70 nucleotides of flanking sequence are required for heat induction (44-47). The 5' extent of the -47 to +30 homologous regions lies within the limits defined by these

deletions. However, except in the comparison between the hsp26 and hsp70 genes, we do not detect homologies in the -65 to -48 sequences of these genes. This region is of interest because nucleotides -66 to -47 in hsp70 have been demonstrated to be essential for heat inducible expression in monkey COS cells (46). Furthermore, a consensus sequence derived from this region and comparable regions of other heat shock genes has been synthesized and appears to promote heat induced transcripts when inserted upstream of a HSV thymidine kinase gene and assayed in COS cells and in Xenopus oocytes (48). Under our search conditions this short 14 nucleotide sequence, which is not well conserved (33-67% match) between the hsp70 standard sequence and other heat shock genes, could not have been detected.

The fact that flanking sequences implicated in the regulation of the heat shock response were not detected by our sequence comparisons or previous comparisons (12,13) may be explained in several ways. First, the various heterologous assay systems used, and the presence of hsp70 genes on plasmid vehicles, may alter the normal functioning of regulatory sequences active in Drosophila chromosomes. Second, these regulatory sequences may not be protein binding sites, but may act instead to induce a particular DNA structure (e.g., Z DNA) and/or chromatin configuration that regulates the response (47). Finally, it is possible that established procaryotic protein-DNA regulatory interactions are not adequate models for these interactions in eucaryotes. Of particular interest in this regard are the sequences shown to be necessary for induction during amino acid starvation of the his 3 and his 4 genes in yeast (49), which are only 8 nucleotides in length.

Eucaryotic promoters appear to consist of two sequence elements. One element is defined by the TATA sequence and directs RNA polymerase II to the correct initiation point in a variety of genes in vivo (50). The second element, usually situated a variable distance upstream from the TATA sequence, appears to modulate the efficiency with which RNA polymerase II initiates transcription in vivo (50). The -66 to -47 region and the more conserved -47 to +30 region of several Drosophila heat shock genes appear to constitute such promoter elements. From the deletion analysis of hsp70 and assay of activity in heterologous systems, it appears that the upstream element functions to enhance promotion by RNA polymerase, since deletions in this region result in low levels of expression during heat shock, instead of constitutive expression (46). In addition, analysis of the in vitro transcriptional activity of mutant globin gene templates has suggested that sequences at the capping-initiation site (-10 to +7) increase the efficiency of transcription

by RNA polymerase II (51). From these results we suggest that the -47 to +30 homologous regions in several Drosophila heat shock genes may function in the proper initiation and perhaps also the efficiency of transcription of these genes during induction. In addition, the sequence disparity of the hps83 gene in these regions relative to the other Drosophila heat shock genes may ultimately be reflected in the differential regulation of this gene at normal and elevated temperatures.

ACKNOWLEDGEMENTS

This research was supported by NIH grant GM 25232. We thank David Finkelstein for bringing to our attention the unusual distribution of charged amino acids in the hsp83 sequence and for communicating the yeast hsp90 sequence prior to publication. We also thank Drs. E. Keller, V. Vogt, and J. de Banzie for comments on the manuscript.

REFERENCES

1. Ashburner, M. and Bonner, J.J. (1979) Cell 17, 241-254.
2. Holmgren, R., Livak, K., Morimoto, R., Freund, R. and Meselson, M. (1979) Cell 18, 1359-1370.
3. Storti, R.V., Scott, M.P., Rich, A. and Pardue, M.L. (1980) Cell 22, 825-834.
4. Lindquist, S. (1980) Develop. Biol. 77, 463-479.
5. Buzin, C.M. and Petersen, N.S. (1982) J. Mol. Biol. 158, 181-201.
6. O'Connor, D. and Lis, J.T. (1981) Nucl. Acids Res. 9, 5075-5092.
7. Chomyn, A. and Mitchell, H.K. (1982) Insect Biochem. 12, 105-114.
8. Bonner, J. Jose (1981) Develop. Biol. 86, 409-418.
9. Graziosi, G., Micali, F., Marzari, F. de Cristini and Savoini, A. (1980) J. Exp. Zool. 214, 141-145.
10. Zimmerman, J.L., Petri, W. and Meselson, M. (1983) Cell 32, 1161-1170.
11. Wu, C. (1980) Nature 286, 854-860.
12. Holmgren, R., Corces, V., Morimoto, R., Blackman, R. and Meselson, M. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 3775-3778.
13. Ingolia, T.D. and Craig, E.A. (1981) Nucl. Acids Res. 9, 1627-1642.
14. Mason, P.J., Torok, I., Kiss, I., Karch, F. and Udvardy, A. (1982) J. Mol. Biol. 156, 21-35.
15. Efstratiadis, A., Vournakis, J.N., Donis-Keller, H., Chaconas, G., Dougall, D.K. and Kafatos, F.C. (1977) Nucl. Acids Res. 4, 4165-4174.
16. Boyer, H.W. and Roulland-Dussoix, D. (1969) J. Mol. Biol. 41, 459-472.
17. Marinus, M.G. and Morris, N.R. (1974) J. Mol. Biol. 85, 309-322.
18. Wensink, P.C., Finnegan, D.J., Donelson, J.E. and Hogness, D.S. (1974) Cell 3, 315-325.
19. Lis, J.T. (1980) Methods Enzymol. 65, 347-353.
20. Smith, D.R. and Calvo, J.M. (1980) Nucl. Acids Res. 8, 2255-2274.
21. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
22. Lis, J.T., Prestidge, L. and Hogness, D.S. (1978) Cell 14, 901-919.
23. Lis, J.T., Ish-Horowicz, D. and Pinchin, S.M. (1981) Nucl. Acids Res. 9, 5297-5310.
24. Guo, L.-H. and Wu, R. (1982) Nucl. Acids Res. 10, 2065-2084.

25. Favaloro, J., Treisman, R. and Kamen, R. (1980) Methods Enzymol. 65, 718-749.
26. Breathnatch, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
27. Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. and Davidson, N. (1982) Cell 29, 1027-1040.
28. Sollner-Webb, B. and Reeder, R.H. (1979) Cell 18, 485-499.
29. Mount, S.M. (1982) Nucl. Acids Res. 10, 459-472.
30. Hentschel, C., Irminger, J.-C., Bucher, P. and Birnstiel, M.L. (1980) Nature 285, 147-151.
31. Busslinger, M., Portmann, R., Irminger, J.-C. and Birnstiel, M.L. (1980) Nucl. Acids Res. 8, 957-977.
32. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corski, P., Kedinger, C. and Chambon, P. (1980) Science 209, 1406-1411.
33. Szalay, A.A., Grohmann, K. and Sinsheimer, R.L. (1977) Nucl. Acids Res. 4, 1569-1578.
34. Kozak, M. (1981) Nucl. Acids Res. 20, 5233-5252.
35. Arrigo, A.P., Fakan, S. and Tissieres, A. (1980) Develop. Biol. 78, 86-103.
36. Barker, W.C. and Dayhoff, M.O. (1982) Proc. Nat. Acad. Sci. U.S.A. 79, 2836-2839.
37. Fristensky, B., Lis, J. and Wu, R. (1982) Nucl. Acids Res. 10, 6451-6463.
38. Little, J.W. and Mount, D.W. (1982) Cell 29, 11-22.
39. Valentin-Hansen, P. (1982) EMBO Journal 1, 1049-1054.
40. Karch, F., Torok, I. and Tissieres, A. (1981) J. Mol. Biol. 148, 219-230.
41. Southgate, R., Ayme, A. and Voellmy, R. (1983) J. Mol. Biol. 165, 35-57.
42. Ish-Horowicz, D. and Pinchin, S.M. (1980) J. Mol. Biol. 142, 231-245.
43. Udvardy, A., Sumegi, J., Toth, E.C., Gausz, J., Gyurkovics, H., Schedl, P. and Ish-Horowicz, D. (1982) J. Mol. Biol. 155, 267-280.
44. Corces, V., Pellicer, A., Axel, R. and Meselson, M. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 7038-7042.
45. Mirault, M.-E., Southgate, R. and Delwart, E. (1982) EMBO Journal 1, 1279-1285.
46. Pelham, H.R.B. (1982) Cell 30, 517-528.
47. Bienz, M. and Pelham, H.R.B. (1982) EMBO Journal 1, 1583-1588.
48. Pelham, H.R.B. and Bienz, M. (1982) EMBO Journal 1, 1473-1477.
49. Donahue, T.F., Daves, R.S., Lucchini, G. and Fink, G.R. (1983) Cell 32, 89-98.
50. Grosschedl, R. and Birnstiel, M.L. (1982) Proc. Nat. Acad. Sci. U.S.A. 79, 297-301.
51. Talkington, C.A. and Leder, P. (1982) Nature 298, 192-195.
52. Hung, M.-C. and Wensink, P.C. (1981) Nucl. Acids Res. 9, 6407-6419.
53. Meyerowitz, E.M. and Hogness, D.S. (1982) Cell 28, 165-176.
54. Goldberg, M. (1979) Dissertation (Stanford University, Stanford, CA).
55. Sanchez, F., Tobin, S.L., Rdest, U., Zulauf, E. and McCarthy, B.J. (1983) J. Mol. Biol. 163, 533-551.