Conserved features of eukaryotic hsp70 genes revealed by comparison with the nucleotide sequence of human hsp70

gene structure/heat shock/evolution

CLAYTON HUNT* and RICHARD I. MORIMOTO†

Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, IL 60201

Communicated by E. Margoliash, June 3, 1985

ABSTRACT We have determined the nucleotide sequence of the human hsp70 gene and 5' flanking region. The hsp70 gene is transcribed as an uninterrupted primary transcript of 2440 nucleotides composed of a 5' noncoding leader sequence of 212 nucleotides, a 3' noncoding region of 342 nucleotides, and a continuous open reading frame of 1986 nucleotides that encodes a protein with predicted molecular mass of 69,800 daltons. Upstream of the 5' terminus are the canonical TATAAA box, the sequence ATTGG that corresponds in the inverted orientation to the CCAAT motif, and the dyad sequence CGGAGA/ATTCCC that shares homology in 12 of 14 positions with the consensus transcription regulatory sequence common to Drosophila heat shock genes. Comparison of the predicted amino acid sequences of human hsp70 with the published sequences of Drosophila hsp70 and Escherichia coli dnaK reveals that human hsp70 is 73% identical to Drosophila hsp70 and 47% identical to E. coli dnaK. Surprisingly, the nucleotide sequences of the human and Drosophila genes are 72% identical and human and E. coli genes are 50% identical, which is more highly conserved than necessary given the degeneracy of the genetic code. The lack of accumulated silent nucleotide substitutions leads us to propose that there may be additional information in the nucleotide sequence of the hsp70 gene or the corresponding mRNA that precludes the maximum divergence allowed in the silent codon positions.

The hsp70 gene in eukaryotic cells is a member of a small family of heat shock or stress-induced genes whose expression is regulated at the transcriptional and translational levels by diverse chemical and biological stresses, including inhibitors of energy metabolism, amino acid analogues, heavy metals, anoxia, and heat shock (1, 2). It has been shown recently that the expression of hsp70 is induced by the viral oncogenes of the DNA tumor viruses adenovirus 5, simian virus 40 (SV40), and polyoma and by the mouse rearranged c-myc oncogene (3–5). These oncogenes share at least two interesting features in that their products are subcellularly localized in the nucleus and are implicated in the immortalization of primary cells in culture (6–11).

Although the function of hsp70 is unknown, we can speculate on its role based on the extensive homology between Drosophila hsp70 and the Escherichia coli heat shock protein dnaK. The amino acid sequence of E. coli dnaK derived from its recently determined nucleotide sequence reveals that it is 48% homologous to Drosophila hsp70 (12). Mutations at the dnaK locus affect normal growth of E. coli (13, 14). Therefore, it may be reasonable to suggest that hsp70 is also involved in eukaryotic cell growth.

hsp70 genes have been isolated from yeast, Drosophila, Xenopus, chicken, mouse, and human genomes (refs. 15–20; unpublished data). In yeast and Drosophila, members of the hsp70 multigene family have been isolated and many of the corresponding genes have been sequenced. Two common features have emerged from these studies: hsp70 is encoded by multiple genes and both Drosophila and yeast genomes encode families of hsp70-related genes. Some members of this gene family are heat shock inducible, whereas others are transcribed at normal growth conditions.

Previously, we have cloned and characterized a human genomic DNA fragment containing a functional hsp70 gene (15). The hsp70 gene is transcribed at low levels in human cells at normal growth temperatures and is induced over 20-fold by a transient heat shock at 43°C. By characterization of the mRNA transcribed in heat-shocked human cells we suggested that the primary transcript of the hsp70 gene is uninterrupted. This is an uncommon feature among vertebrate genes that have been isolated and characterized to date. In this study we present the entire nucleotide sequence of the human hsp70 gene. The hsp70 gene is transcribed as an uninterrupted mRNA that encodes a protein of 69,800 daltons. Comparison of the human hsp70 gene to Drosophila hsp70 and E. coli dnaK genes reveals extensive homologies in both the nucleic acid and predicted amino acid sequences.

MATERIALS AND METHODS

Nucleotide Sequence Analysis. The human hsp70 gene was originally identified and mapped within a 17-kilobase (kb) DNA insert in a recombinant λ phage from the human genomic library (15) kindly provided by T. Maniatis (21). The 2.3-kb BamHI–HindIII and the 2.8-kb BamHI fragments were isolated from the λ phage recombinant H3-1, subcloned into the plasmid vector pAT153, and amplified. Plasmid pH2.3 was digested with HindIII and divided into aliquots that were redigested with either Sca I, Sma I, Cla I, Bgl II, or Pst I. Sma I and Sca I double digests were directionally cloned into the vector M13mp8, which had been digested with Sma I and HindIII. The Bgl II and Pst I digests were ligated into the vector M13mp8, which had been digested previously by BamHI and HindIII or Pst I and HindIII, respectively. The Cla I and HindIII fragments were inserted into Acc I- and HindIII-digested mp8 vector. Double-enzyme digests of pH2.3 with Bgl II and BamHI yielded two insert fragments, which were cloned in both orientations into BamHI-cleaved M13mp9. Xma I digestion of pH2.3 yielded a single insert fragment, which was ligated into Xma I-cut mp9; however, only the 5' → 3' orientation was obtained among the resulting clones. A Bgl II digest of pH2.3 was divided into two aliquots, which were restricted further with either Pst I or Mst I and then cloned into Pst I and BamHI

Abbreviations: kb, kilobase(s); SV40, simian virus 40.
*Permanent address: Dana–Farber Cancer Center, Harvard Medical School, Boston, MA 02115.
†To whom reprint requests should be addressed at: Department of Biochemistry, Molecular and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, IL 60201.
mp8 or Sma I and BamHI mp9, respectively. A final set of hsp70-coding region clones was constructed by digesting pH2.3 with Rsa I and cloning the resulting DNA fragments into Sma I-digested M13mp9. The Rsa I fragments were also digested by Bgl II and subcloned into Sma I and BamHI M13mp9. The 5′ terminus of the hsp70 gene, located in pH2.8, was digested with BamHI, subcloned into the BamHI site of mp8, and sequenced. The nucleotide organization of the entire hsp70-coding region was determined for both DNA strands by using the dideoxy nucleotide chain-terminator method (22).

The primary DNA sequence was compiled and edited on a microcomputer using homology programs to align sequences (23, 24). Amino acid sequences were obtained by computer-assisted analysis of the nucleotide sequence of the human and Drosophila hsp70 genes and the E. coli dnaK gene (12, 18).

**RESULTS**

The organization of the human hsp70 gene and restriction maps of the plasmid subclones pH2.3 and pH2.8 are shown in Fig. 1. The positions of the 5′ and 3′ termini of the hsp70 transcription unit were determined previously by the methods of S1 nuclease protection and reverse transcription with poly(A) mRNA from control and heat-shocked HeLa cells (15). From these results, we concluded that the primary transcript is 2.4 kb and corresponds to a cytoplasmic poly(A) mRNA of 2.6 kb. We had suggested previously that the human hsp70 gene was transcribed as an uninterrupted mRNA. To confirm this observation and to determine the level of homology between the human hsp70 gene and the corresponding genes in Drosophila and E. coli, we obtained the nucleotide sequence of the entire hsp70 gene. The M13 subclones used as templates for DNA sequencing and the strategies of sequence determination are shown in Fig. 1.

The entire nucleotide sequence of the human hsp70 gene and flanking regions was obtained by using the dideoxy chain-terminator method (22) using various restriction fragments subcloned into the vector M13. The nucleotide sequence from -276 to +2391 nucleotides along with the predicted amino acid sequence of the hsp70 gene is shown in Fig. 2.

The 5′ terminus of hsp70 mRNA is located 150 nucleotides upstream of the BamHI site between pH2.8 and pH2.3, as shown in Fig. 1 (15), and is indicated by (+) positioned above an adenine residue. The canonical sequence TATAAA is located at a position -22 to -28 nucleotides upstream of the initiation site of transcription. The 5′ transcribed nontranslated segment is 211 nucleotides, followed by the initiator codon AUG that signals the start of translation for hsp70. Analysis of the nucleotide sequence in all three reading frames revealed only one continuous open reading frame of 1986 nucleotides starting at the first AUG and extending to the in-frame amber termination codon TAG (Figs. 1 and 2). Multiple termination codons are located in reading frames (Fig. 1). The distance between the termination codon TAG and the canonical poly(A) signal AATAAA is 242 nucleotides. This is in excellent agreement with the location of the 3′ terminus of hsp70 mRNA positioned previously by S1 nuclease analysis (15).

The derived amino acid sequence of human hsp70 yields a protein of 69,800 daltons, which agrees well with the value of 70,000 daltons calculated from the electrophoretic mobility in NaDodSO4/polyacrylamide gels (15).

**DISCUSSION**

We have determined the entire nucleotide sequence of a human hsp70 gene. From the analysis of the nucleotide sequence together with our published data on the position of the 5′ and 3′ termini (15), we conclude that the hsp70 gene is transcribed as an uninterrupted primary transcript of ~2400 nucleotides that corresponds to a polyadenylated cytoplasmic mRNA of 2.6 kb. The hsp70 gene sequence contains one continuous open reading frame that encodes a protein of predicted size, 69,800 daltons. Analysis of the nucleotide sequence reveals that the 5′ terminus is flanked by the canonical promoter sequence TATAAA (25) and the 3′ terminus is flanked by the poly(A) signal AATAAA (26).

Although the lack of intervening sequences is very unusual for vertebrate genes, this feature is common among the hsp70

---

**Fig. 1.** Physical organization of the human hsp70 gene and DNA sequencing strategy. (A) Restriction map of plasmid subclones pH2.8 and pH2.3 (15). The locations of the restriction sites shown for pH2.3 are as shown. (B) DNA sequencing strategy. M13mp8 or M13mp9 subclones were derived and subjected to dideoxy chain-terminator DNA sequencing (22). The origin and extent of DNA sequence obtained with each plasmid subclone are shown. (C) Identification of open reading frames. The DNA sequence shown in Fig. 2 was analyzed for the presence of initiator methionine codons (↓) and termination amber or ochre codons (↑). (D) Basic features of the human hsp70 gene. The location of the conserved heat shock transcription element (CTGGAAT/ATTCCG), canonical "TATA" box (TATAAA), positions of the 5′ and 3′ termini of hsp70 mRNA, the initiator (ATG), terminator (TAG), and the poly(A) site (AATAAA) are shown.
genes of Drosophila, yeast, chicken, and mouse (refs. 16–19; unpublished data). It may not be coincidental that hsp70 genes lack intervening sequences. Perhaps this feature is significant for a gene that is rapidly activated at the transcriptional level in a wide range of cell types. Other human genes that lack intervening sequences include histones and α-interferon (27, 28).

We have compared the sequences upstream of the 5' terminus of the human and Drosophila hsp70 genes. The sequence ATTGG at −67 corresponds in the inverted orientation (CATAAT) to the CAT motif found upstream of the transcriptional start site. The dyad sequence CTGGAA/ATTCCCG centered at −100 nucleotides upstream of the 5' terminus corresponds in 12 of 14 positions with the consensus (CTNGAATNTTCTAG) heat shock transcriptional regulatory element of the Drosophila heat shock genes (29–31).
Related sequence elements are also found at a similar position upstream of the mouse and chicken hsp70 gene (unpublished data) and a soybean heat shock gene (32). Although we have not directly demonstrated that this dyd symmetrical element located upstream of the human hsp70 gene is the functional heat shock element, sequences upstream of the 5' terminus are necessary for heat shock-induced transcription (15). Between -115 and -131 we note the sequence GGGCaggacgGAGGC that contains two copies of the pentanucleotide sequence GA/GGGC. This pentanucleotide sequence upstream of the hsp70 gene corresponds by sequence homology to the SV40 or polyoma large T-antigen nucleotide-binding sites involved in the regulation of viral DNA replication, repression of early transcription, and stimulation of late viral transcription (33-36). The position of these pentanucleotide sequences in the hsp70 upstream region is provocative. They could provide an interesting mechanism for the activation of hsp70 gene expression by SV40 and polyoma (4, 5).

The extent of hsp70 conservation among eu-karyotic species is revealed by similarities in the size and charge of hsp70 and conserved antigenic sites. These similarities have been extended by the demonstration that the hsp70 genes from yeast, Drosophila, and E. coli are highly related (12, 18, 19).

Our analysis of the human hsp70 gene allows an evolutionary comparison of the hsp70 genes that extends through higher vertebrates. Comparison of the nucleotide sequences reveals that the E. coli and human hsp70 genes are 50% homologous, whereas the Drosophila and human genes are 72% homologous. These sequence homologies were determined with the Bestfit computer program in which only a few in-frame gaps were inserted to optimize homology (23, 24). Whereas the hsp70 genes are highly related throughout the entire transcriptional unit, the 5' and 3' transcribed nontranslated segments are less conserved than the coding regions. We find considerable homology throughout the entire protein coding region among hsp70 from different species, with the amino and carboxyl termini slightly less related.

The predicted amino acid sequence of human hsp70 shares 73% homology with Drosophila hsp70 and 47% homology with E. coli dnaK (Fig. 3). Although regions of homology extend throughout the entire protein, the carboxyl-terminal region (amino acid residues 550-640) has diverged most extensively. The level of homology between the amino acid sequences of human to Drosophila or E. coli hsp70 was averaged over consecutive 25 amino acid residues for the entire protein (Fig. 4). The 73% homology between human and Drosophila hsp70 corresponds to a range of 40% homology between residues 600 and 625 to 96% homology between residues 150 and 175, with regions from residues 125 to 525 retaining the highest overall conservation. The comparison between human hsp70 and E. coli dnaK reveals two major domains of conserved sequences, from residues 125 to 250 and from residues 375 to 500. Within each of these segments are stretches of 20-40 amino acids with nearly identical sequence. Whereas amino acid residues 250-375 of E. coli dnaK are divergent from both human and Drosophila hsp70, the corresponding region of Drosophila hsp70 is homologous to the human sequence. However, these changes in the amino acid sequence between E. coli dnaK and Drosophila and human hsp70 are conservative substitutions such that the structural domains revealed by the hydrophathy index are maintained (data not shown). We suggest that the extensive homologies among eu-karyotic hsp70 proteins and the divergence in the central region of the hsp70 protein with E. coli dnaK may define unique functional requirements. It is of interest to note that the carboxyl-terminal segment of human and Drosophila hsp70 and E. coli dnaK have diverged extensively from each other. Despite the larger number of changes in amino acid sequence from residue 550 to the carboxyl terminus, 17 of 100 residues remain conserved among the three proteins, with the insertion of only two gaps for alignment. Of particular notice are the three amino acids Glu-Glu-Val found at the termini of all three hsp70 proteins.

The extensive homologies shared among E. coli, Drosophila, and human hsp70 suggest possible functions for hsp70 (12-14). The synthesis of E. coli dnaK protein is induced by heat shock. The function of dnaK is associated with host DNA synthesis and is essential for replication of bacteriophage λ (13). The phenotype of mutations at the dnaK locus
Biochemistry: Hunt and Morimoto

![Graph showing homology to human hsp70](image)

**Fig. 4.** Homology plot of conserved amino acids between human and Drosophila (—) and between human and E. coli (——). The level of homology between consecutive stretches of 25 amino acid residues was determined as an average and plotted against the linear protein. Residue 1 corresponds to the amino terminus of hsp70.

is categorized as slow stop in cell growth, a phenotype that is consistent with a role in cell division (13). In human tissue culture cells the transcription of the hsp70 gene and accumulation of hsp70 mRNA is serum stimulated in a growth-dependent manner. Maximum levels of hsp70 mRNA occur between 10 and 18 h after stimulation parallel to the peak of DNA synthesis (37).

The rate of evolution of hsp70 can be approximated from the percent change in the derived hsp70 amino acid sequence from different species divided by the time (in millions of years) since they shared a common ancestor. The rate of evolution is presented as a unit of evolutionary period (UEP) and corresponds to the average time (in millions of years) required for a 1% difference in amino acid sequence of a particular protein to arise between two evolutionarily distinct lineages. A value for UEP can be calculated for hsp70 assuming that Drosophila and humans shared a common ancestor 6 x 10^8 years ago (38, 39). The Drosophila and human hsp70 proteins differ in 27% of the amino acid positions from which a UEP value of 22 is calculated. A similar UEP value is obtained from comparison of chicken or mouse hsp70 to Drosophila and human hsp70 (unpublished data). The UEP for hsp70 reveals a level of conservation similar to cytochrome c (UEP = 20) and considerably less than observed for α-tubulin (UEP = 550) (38, 40).

The degeneracy of the genetic code allows for changes in the nucleotide sequence that do not affect the amino acid sequence of the protein due to silent substitutions. We have compared the coding regions of the Drosophila and human hsp70 (Fig. 2) and find that the accumulation of silent substitutions for hsp70 is in contrast to other highly conserved proteins such as cytochrome c or tubulin (38–40). There is a constraint on silent substitutions for hsp70 such that only 45% of the available third positions have diverged (Fig. 2) This result provides a logical explanation for the unusually high level of sequence conservation of the hsp70 gene between Drosophila and humans. Because the levels of homology for the amino acid and nucleotide sequences are similar, we suggest that either the hsp70 gene or mRNA contains additional information that selects against the maximum divergence allowed by the conserved amino acids. Upon closer examination of the Drosophila and human nucleotide sequences we find that the codons in which third-position substitutions have not occurred are clustered. Within the amino-coding region of hsp70 mRNA there are at least eight regions of 15–40 nucleotides separated by sequences in which many of the third positions have undergone nucleotide substitutions (Fig. 2). The location of the nucleotide sequences conserved between human and Drosophila hsp70 does not appear to be random. Perhaps these conserved regions have functional significance for the chromatin structure of the hsp70 gene or, more likely, in the secondary structure of the mRNA. The evolutionary constraint in the nucleotide sequence may conserve mRNA structures necessary for translational control of protein synthesis.

We thank B. Wu for discussion throughout this study and comments on this manuscript and R. Kamen for discussions that revealed the presence of SV40 and polyoma T-antigen nucleotide-binding sites upstream of the hsp70 gene. L. Berg provided assistance in computer analysis of sequence data and K. Olson cheerfully prepared this manuscript. This study was supported by grants from the Camille and Henry Dreyfus Foundation, American Cancer Society, Leukemia Foundation, and the National Institutes of Health.