Nucleic Acids Research, 1995, Vol. 23, No. 4 541–549

5′ untranslated leader sequences of eukaryotic mRNAs encoding heat shock induced proteins

Chandrashekhar P. Joshi and Henry T. Nguyen*

Plant Molecular Genetics Laboratory, Mail Stop 2122, Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX 79409-2122, USA

Received June 13, 1994; Revised and Accepted December 23, 1994

ABSTRACT

5′ untranslated leaders (5′ UTLs) are suggested to play a crucial role in the selective translation of their eukaryotic mRNAs encoding heat shock proteins (HSP) during heat stress conditions. However, the structural features of the HSP mRNAs which cause this effect are mostly unknown. We have compiled the 5′ UTLs from about 140 eukaryotic HSP mRNAs including vertebrates, invertebrates, higher and lower plants. A detailed analysis of these sequences according to length, A+T content, context of functional ATGs and presence of upstream non-functional ATGs was made. We observed that all these features were similar to the earlier studies in the literature based on data from HSP as well as non-HSP mRNAs. These observations were reconfirmed by intra-specific comparison of 5′ UTLs from HSP and non-HSP genes. Similar to the translation element involved in the selective translation of mRNAs in polioviruses, a search for a short sequence motif complementary to highly conserved 18S rRNA was performed using a HSP mRNA database. The majority of the HSP mRNA sequences (77%) contained one or more small sequence motifs suggesting that they may function as internal ribosome entry sites for selective initiation of translation during heat stress.

INTRODUCTION

Synthesis of heat shock proteins (HSPs) in response to increases in the normal growth temperature of an organism is a universal phenomenon (1–3). HSPs are also produced in a variety of stress conditions other than heat stress indicating that they could be common stress proteins (4). In addition to their proposed role as cell protectors during stress, they also provide an excellent system to study transcriptional and translational regulation of gene expression. Two important changes in patterns of gene regulation occur in a cell in response to heat shock. First, there is an almost exclusive synthesis of HSP mRNAs and second, these HSP mRNAs are preferentially translated while normal cellular mRNAs (non-HSPs) are translationally suppressed (1,5). It has been suggested that this discriminatory behavior of a cell is due to a sudden demand for protective proteins. The 5′ untranslated leader (5′ UTL) of HSP mRNAs is suggested to play a key regulatory role in this process (1).

In *Drosophila*, 5′ UTLs of HSP70 and HSP22 are necessary for their preferential translation during heat shock (6,7). In the HSP70 leader, two regions, one in the middle of the leader and the other at the 5′ end, were highly conserved and were thought to be important for preferential translation. Deletion of both sequences resulted in the loss of preferential translation. However, their function could be maintained if at least one sequence motif was present, suggesting the existence of a redundancy in the function of these sequence motifs. Although the precise sequence motif within the leader was not identified, it was suggested by McGarry and Lindquist that it must be present somewhere in the leader (7). Klemenz *et al.*, (8) have also demonstrated the need of HSP70 5′ UTL for the preferential translation of normal Adh mRNA synthesized during heat stress periods under the control of the HSP70 promoter. Recently, Pitto *et al.*, (9) examined the role of the HSP70 leader from maize that conferred on a reporter mRNA the ability to escape from thermal repression of translation in heat shocked plant protoplasts. Gallie and Walbot (10) demonstrated that the tobacco mosaic virus leader sequence behaves similar to a heat shock protein leader during heat shock. Using deletion analysis, they reported a functional redundancy between two different highly conserved motifs, namely, an 8 bp direct repeat and a 25 bp $(CAA)_n$ region. These studies indicated the importance of 5′ UTLs in HSP mRNAs for their preferential translation during heat shock. However, eukaryotic HSP mRNA leader sequences have not been examined for the existence of any conserved nucleotide sequence motifs that might suggest a common regulatory function. Generally, HSP mRNAs are described as having unusually long leaders that are adenine rich, have little secondary structure and have conserved sequences in the middle and at the 5′ ends (11). This description fits several *Drosophila* HSP genes but is not universal. One of our objectives in the present studies was to characterize the occurrence of these features in eukaryotic HSP mRNA leaders.

The 5′ UTL sequences of the mature eukaryotic mRNAs are implied to play a crucial role in initiation and efficiency of translation (12). A model of translational initiation was proposed that includes binding of the 43S initiation complex at the capped 5′-end of the mRNA and linear scanning of the entire 5′ UTL until the first ATG codon in good context is found (the nucleotide sequences in this paper are expressed as the plus strand of DNA

* To whom correspondence should be addressed

from which mRNA sequence can be derived by replacing T by U). At this point, the 60S subunit of ribosome joins the 40S subunit and translation begins. The optimum context of ATG codon has been proved to be a purine residue at the –3 position by mutational analysis (12). Other positions in the context of a functional ATG are known to vary among eukaryotic species, for example, CC(A/G)CC in vertebrates (13), AAACA in plants (14), (A/T)A(A/C)A(A/C)A in yeast (15) and (C/A)A(A/C)A in *Drosophila* (16). Recently, these observations have been further extended by collecting and analyzing larger and more representative data sets (17). These observations suggest that in various eukaryotic species, positions other than –3 are not highly conserved. Although 'the first ATG as a translation initiation codon' is common in eukaryotes, there are a few exceptions to this rule (12). In such cases, either upstream ATGs have a suboptimum context (no purine at –3 position) and are followed by in-frame stop codons or the ribosomes simply bypass such ATG codons without initiating translation. Translation in such cases reinitiates at the downstream ATG with optimum context (12). Whether a special group of mRNAs, e.g. HSP mRNAs, have a deviant context as compared to non-HSP mRNAs has never been studied, although such a possibility was indicated earlier (14). This aspect is examined in this paper.

The translational machinery of eukaryotes is highly complex, consisting of several initiation factors, elongation factors and release factors (18). The overall process of translation begins when the cap of mRNA is recognized by eukaryotic initiation factor complex eIF-4F which facilitates the melting of the mRNA 5′ end secondary structure and 43S ternary complex binding to mRNA. This eIF-4F complex consists of a 25 kDa cap recognition protein (eIF-4E), eIF-4A and a 220 kDa protein of unknown function. The cytoplasmic concentration of eIF-4E is low and mRNAs compete for binding to the eIF-4F complex (19). This competition could be the most important event affecting the process of translation.

Recently, new insights have been gained into the mechanisms of preferential translation initiation of HSP mRNAs. Lamphear and Panniers (20) reported that heat shock impairs the recognition of the 5′ cap by the eIF-4F complex and shows reduced phosphorylation of eIF-4E which in turn inactivates the function of these initiation factor complexes. This and several other studies indicated that heat shock interferes with the maintenance of the initiation factors and that translation of HSP mRNA is less cap-dependent. Joshi-Barve *et al.* (21) have shown that the use of antisense RNA directed against eIF-4E reduces the level of eIF-4E in HeLa cells. However, certain proteins continued to be synthesized in these cells and were shown to be HSP90, HSP70, HSP60 and HSP27. This experiment clearly demonstrates that HSP mRNAs have little or no requirement for cap-recognition machinery and that these mRNAs may use an alternative, cap-independent mechanism for translation initiation. Thus, non-availability of active eIF-4F appears to regulate gene expression of non-HSP mRNAs during heat shock and translation initiation of HSP mRNA is less cap recognition-dependent than non-HSP mRNAs.

A similar situation is found in animal cells infected with polioviruses or other picornaviruses (22) where ribosomes bind at the internal ribosome entry site (IRES) to uncapped poliovirus mRNA in a cap-independent fashion (23). Destruction of translation initiation factors such as eIF-4F from host cells is one of the strategies this group of viruses employs to stop host protein synthesis and continue viral translation. Recently, Pilipenko *et al.* (24) identified a smaller motif, UUUCC, and an appropriately spaced ATG in poliovirus RNAs as some of the possible elements of polio virus IRES. This is similar to the prokaryotic translation where the Shine–Dalgarno sequence is important. These two sequences in polioviruses are partially complementary to highly conserved 18S rRNAs which are the major constituents of 43S ribosomal subunits. This indicates that a mechanism similar to the picorna viral translation may exist in normal eukaryotic cells where discrimination in translation of a particular kind of mRNA is common and this virus has simply used the host's existing mechanism. More biochemical and genetic evidence is required to support the model that 18S rRNA is intimately involved in recognition of IRES. Although the internal initiation of translation of HSP mRNAs during heat shock was a possibility, the real proof was provided by Macejak and Sarnow (25) who inserted the 5′ UTL of heat shock protein, BiP (immunoglobulin heavy chain binding protein) or GRP 78 (78 kDa glucose regulated protein) between two cistrons, and translation was shown to be initiated internally. No nucleotide sequence similarity was, however, noted between BiP leader and picorna virus leaders. This raises the possibility that HSP mRNAs may also have a smaller IRES motif. We have examined such a possibility by compilation and analysis of most of the leader sequences from eukaryotic mRNAs encoding HSP mRNAs. This is likely to provide some insight into the mechanism of internal initiation of HSP mRNAs, assuming this mechanism is really common in the eukaryotic heat shock response. In addition, we also used this compilation to answer the following questions regarding 5′ UTL sequences of eukaryotic HSP mRNAs: (i) What are the lengths of these leaders? (ii) What are their nucleotide compositions? (iii) How many of these leaders have upstream ATGs and in what contexts? (iv) Is there any specialty of the context of the translation initiation codon? (v) Is there a short sequence motif complementary to highly conserved 18S rRNA which could act as IRES for HSP mRNAs?

Analysis of 5′ UTLs from HSP genes will not be complete unless a comparison of these data is made with non-HSP genes. Unfortunately, it is difficult to define the term non-HSP genes that encompasses a heterogeneous group of genes that are not HSPs and these mRNAs may be under different translational control. There are excellent compilation and analysis papers in literature dealing with 5′ UTLs from vertebrates (13) and higher plants (14) where less than 5% of the genes included in the compilation were from HSP group and we have used these data in this paper as representative of non-HSP genes. Similar compilation and analysis papers are also available for functional ATG context of Drosophila genes (16) and yeast genes (15) but details regarding 5′ UTL lengths and their AT content are not available for these two groups. We have, therefore, randomly selected 10 non-HSP genes each from four groups of eukaryotes namely, vertebrates, invertebrates, higher plants and lower plants and compared their structural features with HSP genes that are included in this compilation.

METHODS

Selection of HSP gene sequences for the present compilation

The GenBank (Release 76.0, 1993) and EMBL (Release 34.0, 1993) databases were used for this compilation. Seven database

sections namely, invertebrate, primate, rodent, other mammals, other vertebrates, plants and unannotated were searched using the key word 'heat shock'. All sequences having this keyword in their definition were retrieved from the databases and checked to see if they met the following criteria:

Context of translation start codon. Sequences were included in the present compilation if 12 bases upstream and three bases downstream of the proposed translation start codon were present.

5′ UTL sequences. Eukaryotic cDNAs encoding HSPs were selected for this study if the leader sequences (5′ untranslated regions) were longer than 10 bp. Genomic sequences encoding HSPs were included if the transcription start site was mapped by primer extension or S_1 nuclease analysis and this feature was mentioned in the feature table of the sequence in the databanks. In the absence of these data, an origin, 30 bp downstream of the proposed TATA box was regarded as a putative transcription start site. This is the approximate distance between these two motifs (14,26). One gene sequence/gene family/eukaryotic species was included in the compilation without any reference to their evolutionary relationships.

All the sequences truncated in the above mentioned regions or clearly marked as pseudo-genes were excluded. Identical sequences encoding the same proteins from the same species were also deleted.

Data analysis

Context of translation start site. All selected sequences were aligned with 12 bases upstream and three bases downstream from the proposed ATG codon. Consensus sequences were determined separately for vertebrates, invertebrates, higher plants and lower plants using the criteria described by Cavener (16) and used earlier by Joshi (27). A single base was given a consensus status and indicated by a capital letter if the relative frequency of a single nucleotide at a certain position was >50% and greater than twice the relative frequency of the second most frequent nucleotide. When no single base satisfied these criteria, a pair of bases were assigned co-consensus status if the sum of the relative frequencies of the two nucleotides exceeded 75%. If neither of these two criteria was fulfilled at a position, it was denoted by the most frequent nucleotide in lower case.

5′ UTLs. Using the computer program routine 'lineup' from the GCG package [Genetics Computer Group, WI (28)], all HSP leaders were collected according to the molecular weight of HSPs they encode or by taxonomic groups. These data were used for the calculation of A, T, G and C compositions and the lengths of the leader sequences. Further, the routine 'pileup' was employed to align these sequences according to their homologies and a search for common sequence motifs was made in these data sets. Finally, 5 bp motifs from conserved 3′ ends of the 18S rRNAs were searched against these data and common sequences occurring in most of the leaders were identified.

Selection of non-HSP genes for the present compilation

Ten non-HSP genes from one species each of four representative groups of eukaryotes namely, vertebrates (e.g. mouse), invertebrates (e.g. *Drosophila*), higher plants (e.g. *Arabidopsis*) and lower plants (e.g yeast) were randomly selected from GenBank/EMBL database. The selection criteria for the compilation of genes were similar to HSP gene collection as described above. 5′ UTLs from these genes were further used for their lengths, AT content, consensus sequence context of functional ATG, presence of upstream ATG and presence of short 5 bp motifs similar to HSP genes. Ten or more HSP genes from these four species (except in case of mouse where three more rat genes were also included) were again analyzed for the same attributes and data between HSP and non-HSP genes from same group were used for intra-specific comparison.

RESULTS AND DISCUSSION

This is the first compilation where 5′ UTLs from the majority of available eukaryotic HSP mRNAs have been collected and analyzed. Previous collections were restricted to less than 10 sequences and included sequences from a single organism, *Drosophila* (6,11). The present collection includes plant and animal sequences, providing a general picture of eukaryotic HSP leaders. Although some HSP leaders are known to confer preferential translation on the downstream coding region, what *cis*-acting elements are or are likely to be involved in this process is unknown and a collection of sequences as presented here is expected to provide a basis for this and further investigations.

Searching through over 65 000 nucleic acid sequences from GenBank/EMBL databanks using the key word 'heat shock' resulted in more than 300 entries. Manual screening of the feature tables and use of selection criteria described in the Methods section indicated that about half the entries belonged to one of the following categories and were not considered for the present compilation: (i) Identical sequences reported from two or more different laboratories. (ii) Feature tables lacking the necessary data for the compilation. (iii) cDNAs encoding the same proteins but with different 5′ ends (in this case, the longest cDNA leaders were included in the compilation). As a result, 166 sequences were included for the comparison of the context sequences for proposed functional ATG codon and 139 sequences yielded information regarding 5′ UTLs of HSP mRNAs (Table 1).

Eukaryotic HSPs are generally classified either as in a relatively high molecular weight group (in the range of 60–104 kDa) or as in a low molecular weight group (in the range of 8–27 kDa). High molecular weight (HMW) HSPs are further classified into subgroups, such as HSP90, HSP70 and HSP60 families (1,2). Some genes encoding HMW HSPs are designated as heat shock cognate genes (HSC genes) because they are also expressed under non-stressed conditions and are not significantly induced by heat stress. These genes are included in the present compilation with the assumption that they might be under the same translational control as HSPs due to their translation during heat shock. Proteins of HSP60 family are generally components of chloroplasts and mitochondria. While animal cells do produce some low molecular weight (LMW) HSPs, higher plant cells produce an abundant amount of LMW HSPs belonging to several gene families ranging from 15 to 27 kDa (5).

The most interesting animal HSP genes included in the present compilation are genomic sequences of the HSP90 family which have an intron in the 5′ untranslated regions of their pre-mRNAs. In fact, small 5′ UTLs (~100 bp) are followed by an intron that is longer than 1000 bp. The next base after the intron is 'A' of the translation initiation codon, ATG. Two members of HSC70 family from human and *Drosophila* also share similar features

Table 1. Data regarding 5′ untranslated leaders of eukaryotic HSP genes

a	b	c	d	e	f	g	h
Vertebrates							
Bos(8)							
X53827	BTHSCP	HSC70	C	34	41	-	GCTTTTGCAACCATGTCT
Cricetulus							
M17169	CRUGRP	GRP78	C	150	34	-	AGCGCCGGCAAGATGAAG
M34561	CRUHSP70A	HSC70	C	49	39	-	CAGCAAGCAACCATGTCT
X51747	CLSHSP	HSP27	C	17	35	-	ACAGCCAAGAACATGACC
Gallus							
M14772	CHKHSP108	HSP108	C	108	32	1	GCATCATGAAGTATGAAG
X15028	GGHSP905	HSP90	G	82	38	-	CCCGCTGCCAAGATGCCG
J02579	CHKHSP	HSP70	G	112	38	-	GAATCTATCATCATGTCT
X57157	GGHSP47	HSP47	C	57	35	-	GCTGCCAAAACCATGCAG
Homo							
J04988	HUMHSP90B	HSP90	G	99	52	-	TTTCTTTTCAAGATGCCT
M30626(8)	HUMHSP86KD	HSP86	G	?	?	?	GTGTCGTTCCAGATGCCT
X07270	HSHSP86	HSP86	C	49	33	-	CACTTAGCCAAGATGCCT
M19645	HUMGRP78	GRP78	G	221	37	-	CTGGCTGGCAAGATGAAG
M12119	HUMDHSP	HSP70	G	?	?	?	CATGCCAGCACCATGTCG
M15432	HUMHSP70D	HSP70	G	215	38	-	GCAGGGAACCGCATGGCC
L08069	HUMDNAJHOM	HSP70	C	82	30	-	CGGCAGTAGAAGATGGTG
M11236	HUMHSP70	HSP70	G	119	30	-	GAAGCTTCAGCCATGCAG
J04505	HUMHSP70E	HSP70e	G	?	?	?	CAGGGAACCGCATGGCC
Y00371	HSHSC70	HSC70	G	83	45	-	ACCCCAGCAACCATGTCC
X03900	HSHSP27	HSP27	G	91	42	-	GAGTCAGCCAGCATGACC
M64673	HUMHSF1	HSF1	C	160	23	2	TCCTTGCTCGAGATGGAT
M65217	HUMHSF2	HSF2	C	88	36	-	GCCGCGTTAACAATGAAG
Mus							
J04633	MUSHSP86A	HSP86	C	260	50	6	CATTCAGCCACGATGCCT
M18186	MUSHSP84A	HSP84	C	94	34	-	TTTCCCGTCAAGATGCCT
M19141	MUSHSPCA	HSC72	C	65	37	-	ACACAAGCAACCATGTCT
M76613	MUSHP7A2	HSP70	G	226	36	-	CGAGCCGGCGCCATGGCC
X55023	MMHSP65R	HSP65	C	18	22	-	CGCCCCGCAGAAAATGCTT
X61753	MMHSF1	HSF1	C	139	32	1	CTTTCGTCCGAGATGGAT
X61754	MMHSF2	HSF2	C	158	34	1	ACCGGGCTAACAATGAAG
Rattus							
M14866	RATGRP78	GRP78	G	206	40	-	AGCGCCGGCAAGATGAAG
X15705	RNHST70A	HSP70	G	112	41	-	TTTTCAGTCAGGATGTCC
M86389	RATHSP27A	HSP27	C	43	42	-	TCAGCCAAGACCATGACC
Salmo							
K02550	SMOHSP70B	HSP70	C	62	58	-	TTATTCGGTAACATGTCT
Sus							
X68213	SSHSP70	HSP70	C	105	49	-	GAAGCCACAGCCATGTCC
Xenopus							
M11915	XLHSP70	HSP70	G	125	54	-	AGGAGCGCAAATATGGCA
X02511	XLHSP30A	HSP30a	G	237	49	2	GTGGAGAGGAGAATGCAG
X02512	XLHSP30B	HSP30b	G	213	52	2	TTGGAGAGGAGAATGCAG
X57963	XLHSP30DD	HSP30d	G	87	60	-	AATCAGAAGAAGATGTTT
X57964	XLHSP30E	HSP30E	G	87	57	-	AATCAGCAGAAGATGTTT
Invertebrates							
Acanthocheilonema							
X68667	AVAV25A	HSP25	C	20	65	-	GTGAACTCAACGATGTCG
Anopheles							
M96662	MSQHSP70A	HSP70	G	232	61	-	AACAAACACAGAATGCCG
M96661	MSQHSP70B	HSP70	G	184	61	-	CAGAGACACAGAATGCCT
Aplysia							
Z15041	ACBIP	GRP78	C	131	59	-	GGTAAAATCAAAATGGAT
Brugia							
M68933	BRPHSPAA	HSP70	G	174	50	1	AATACAGCAAATATGTCA
Caenorhabditis							
M18540	CELHSP70	HSP70	G	?	?	?	TTTACAGTAAAAATGAGT
M26906	CELHSP3B	HSP70	G	?	?	?	CAAACACACAGTATGAAG
K03273	CELHSP1640	HSP16-48	G	51	67	-	TAAACTTCAAGAATGCTC
K03273	CELHSP16C	HSP16-1	G	42	71	-	CAAACTACAATCATGTCA
M14334	CELHSP16D	HSP16-2	G	55	61	-	CAAACTATAATCATGTCA
Drosophila							
K01605	DROHSP83A	HSP83	G	149	68	-	CATACATACAACATGCCA
X03812	DPHSP82	HSP82	G	136	69	-	AACACATACAAGATGCCC
X03813	DVHSP82	HSP82	G	150	67	-	AAGACATACAAGATGCCT
X03811	DSHSP82	HSP82	G	149	66	-	CATAAATACAAGATGCCA
L01498	DROHSC3A	HSC72	C	190	70	-	GAGATCTTCAATATGAAG
L01502	DROHSC5A	HSC71	C	244	56	-	TAAAAAGCCAAAATGCTG
L01501	DROHSC1A	HSC70	C	256	52	-	GCCACAGCCAAGATGCCC
J01085	DROHSC7A1	HSC70	G	?	?	?	GCCACAGCCGACATGCCC
J01105	DROHSP7D1	HSP70	G	242	64	-	AACTCACACACAATGGCT
V00213	DMHSP1	HSP70	G	247	64	-	AACTCACACACAATGCCT
V00218	DMHSP6	HSP27	G	121	70	-	AAAAAATCAAAAATGTCA
V00211	DMHSP26G	HSP26	G	179	68	-	GGAAACGTAAAAATGTCG
J01093	DROHSP673	HSP23	G	114	69	-	TAAAAAACAAAAATGGCA
V00216	DMHSP4	HSP22	G	253	70	-	TTATCAACTACAATGCGT
V00219	DMHSP7	alpha	G	80	60	-	AGTGTTGAAAATATGTAT
X16157	DPHSROM	Omega	G	40	63	-	GTCGCAGTGCCTATGGAA
M26267	DROHSP1	HSP1	G	93	55	-	AAGAAAGTGAAAATGTCG
X07311	DMHGSG2	HSP2	G	61	51	-	CCCATTACTACAATGGCC
X06542	DMHSPG3	HSP3	G	168	55	-	AAGAAAAGTAAAATGCCA
M36114	DROHSC4A2	HSC4	G	>25	60	-	AATTCTTTCCAGATGTCT
M30668	DROHSPHEX	HSF	C	228	62	-	GCTGGTCACTTTATGTCC
Echinococcus							
M63604	ECCGRP	GRP78	C	58	48	-	TATAGTAACGTGATGGGT
Giardia							
X16738	GLHSP70	HSP70	G	?	?	?	AGTATTAGTAGGATGCCT
Hydra							
M84019	HYDHSP701B	HSP70	G	58	72	-	AAAATACAAAAAATGTCT
Leishmania							
X52314	LDHSP70	HSP70	G	?	?	?	ACTGCCGCAGAGATGACA
Onchocerca							
X68668	OVOV251A	HSP25-1	C	21	71	-	GTAAATCCAACAATGTCA
Paracentrotus							
X16544	PLHSP70	HSP70	G	199	64	2	CTGAAAAGCAAAATGGCT
Plasmodium							
M90978	PFAHSP70B	HSP70	G	?	?	?	TGAGTTTTCAAAATGGCC
Schistosoma							
L02415	SCMHSP70X	HSP70	G	55	64	-	ACATATCTTGGAATGCCT
Theileria							
M57386	THEHSP90	HSP90	C	129	74	-	ATCTAATTAACGATGACA
J04653	THE70HSP	HSP70	G	225	72	-	ATAGATTTAAAGATGACA
Trypanosoma							
M15346	TRMHSC	HSC86	G	?	?	?	AAACTCGCAAAGATGACC
M14697	TRBHSP70A	HSP70	C	448	59	5	CCTCTTTGAAGGATGACA
M32140	TRBHSP70US	HSP70	G	?	?	?	AGGATAATAGACATGACG

a	b	c	d	e	f	g	h
Higher plants							
Arabidopsis							
D00710	ATHHSP81	HSP81	G	106	58	2	GCGGATGTTCAGATGGCT
M23105	ATHHSP701	HSC70	G	111	60	-	CTTCCGATAAAAATGTCG
M23106	ATHHSP703	HSC70	G	72	71	-	CTTCGTGTGATAATGGCT
M23107	ATHHSP70A	HSC70	G	85	65	-	GTAGCGTTAACAATGGCT
Z11547	ATCHHSP60	HSP60	C	74	57	-	TTCTCACTCACCATGTAT
X54102	ATHSP21	HSP21	C	76	66	-	TACTCGAAACAAATGGCT
X17295	ATHSP182	HSP18.2	G	42	60	-	AAGCAACGAACAATGTCT
X16076	ATHSP176	HSP17.6	C	48	58	-	GGAAGTGAAACGATGTCT
X63443	ATHSP176II	HSP17.6II	C	38	66	-	ACACAACTAACAATGGAT
X17293	ATHSP174	HSP17.4	G	55	60	-	GTAAGCGAAACGATGTCT
Chenopodium							
S51532	S51532	HSP23.3	C	55	65	-	GTCACTTCAATCATGGCA
S51533	S51533	HSP18.3	C	77	65	-	TCATCATCAACAATGTCG
Daucus							
X60088	DCHSP70	HSP70	G	66	68	-	TAACTGATCACGATGGCT
X53852	DCHSP179	HSP17.9	G	102	70	-	TCATATTAGCAAATGTCG
X53851	DCHSP177	HSP17.7	G	62	69	-	CTACTTCAGCAAATGTCG
Glycine							
X62799	GMHSP70	HSP70	G	106	69	-	TGTGATTGATCAATGGCG
M20363	SOYHSP	HSP26	G	203	66	2	ACAATACAAACAATGGCA
X63198	GMHSP22L	HSP22L	G	?	?	?	TTAGTACTAGAAATGAGG
X07160	GMHSP185	HSP18.5C	G	76	61	-	CCAAGAGATAAAATGTCT
X07159	GMHSP179	HSP17.9D	G	72	69	-	TTTCGCAAAGACATGGAT
M11317	SOYHSP176	HSP17.6L	G	96	70	-	TAACATCAAAGATGTCA
M11318	SOYHSP175	HSP17.5M	G	93	68	-	ATATATCAAAGATGTCT
M11395	SOYHSPGM	HSP17.5E	G	82	65	-	AAAAACCAAAAGATGTCT
X01104	GMHSP2	HSP17.3B	G	103	65	-	TGATAAGAGAAAATGTCT
Helianthus							
X59701	HA176HSP	HSP17.6	C	12	67	-	AAAACATTCACCATGTCA
Lupinus							
X51765	LPHSP70	HSP70	C	71	48	1	GCTGGAGGTGTCATGACT
Lycopersicon							
M96549	TOMHSC80P	HSC80	G	73	60	-	GATCTACAAAAAATGTCG
L08830	TOMBIPGRBC	GRP78	C	119	66	-	AAGGAAGGAGCTATGGCT
X54029	LEHSC170	HSC70	C	37	62	-	CAGTGTAAAGCAATGGCC
X56138	LEHSPIR	HSPLMM	C	43	68	-	AGTTCACTGAAAATGTCT
Medicago							
X58711	MSHSP182	HSP18.2	C	41	67	-	ATTGAAGCGAAGATGTCA
Oryza							
M80186	RICLMMHSP	HSP17.3	C	81	51	-	GCTATTCCGACGATGTCG
M80938	RICHSEA	HSP16.9A	G	131	52	-	GCTAGAGCAACCATGTCG
Petunia							
X13301	PHHSP70R	HSP70	G	?	?	?	TGATTGAGAATAATGGCT
X54103	PHHSP21	HSP22	C	60	75	-	ACTAGTTTTCCAATGGCT
Pharbitis							
M99431	PHNHSP83A	HSP83	G	111	54	-	GTTCTGTGATCAATGGCG
M99430	PHNHSPB	HSP18.4	G	?	?	?	TTACCTAAATAAATGGAT
M99429	PHNSHSPA	HSP17	G	63	63	-	ACTGTTGAAGAAATGGAT
Phaseolus							
X66874	PV70HSP	HSP70	C	130	59	-	TCCATCCCGGCTATGGCT
Pisum							
L03299	PEA70HSP	HSP70	C	45	60	-	CACTCTTCACTCATGGCT
X54739	PSPHSP1	HSP70	C	26	50	-	CTTCGATCCACCATGGCC
M33898	PEAHSP227A	HSP22.7	C	43	65	-	ACATTTCTAAGTATGAGT
M33901	PEAHSP177A	HSP17.7	C	64	63	1	GATTTCAGGCTAATGGAT
Solanum							
Z11984	STHSP70D3	HSP70	G	?	?	?	TTTTCCGTCGTCATGGCC
Z11985	STHSP70D7	HSP70	G	?	?	?	AAACGAAAAGATATGGCC
Z11982	STHSP70DK	HSP70	G	?	?	?	TGAATCTGTAGTATGGCT
Spinacia							
X61491	SOSCE70	HSP70	C	377	60	2	AGGAAGATTATAATGGCC
M87646	SPICPN10X	HSP10	C	49	53	-	CAGTTAGTTGGAATGGCT
Triticum							
X58280	TAHSP26	HSP26.6A	C	90	53	-	TTTCAAAGTGCAATGGCT
X67328	TAHSP266	HSP26.6B	C	82	38	-	GTATCTGGTGCAATGGCC
X58279	TAHSP173	HSP17.3	C	47	32	-	CCTGGCCCGAGAATGGCG
X13431	TAHSPLW	HSP16.9A	C	75	52	-	CAAACACTGACGATGTCG
X64618	TAHSP169B	HSP16.9B	C	15	47	-	ATCAACACCACGATGTCG
Zea							
X03658	ZMHSP701	HSP70	G	>80	49	-	ATTCTGGTGGCAATGGCG
X03714	ZMHSP70I	HSP70	G	107	53	-	CTAGAGGAAGMGATGGCA
Z12114	ZMCPNAGA	HSP60	C	187	37	-	CCTTCCGCAGTCATGTAC
X54076	ZMHSP18K1	HSP18.9	C	239	47	1	GCAGAGCGAGMGATGGAC
X54075	ZMHSP18K2	HSP18.3	C	111	41	-	GACATCCGAGMGATGGAC
X65725	ZMHSP172	HSP17.2	C	78	49	-	CTAGCGTCGACAATGTCG
Lower plants							
FUNGI							
Bremia							
M27825	BRMHSP70	HSP70	G	60	70	-	AATTGAAACACAATGGCT
Kluyveromyces							
X59963	KMHSDNA	HSP	G	?	?	?	CATATTCCAACAATGGCT
X55149	KLHSFG	HSTF	G	?	?	?	AGCCTTTAGATTATGGGT
Neurospora							
M55672	NEUHSP30	HSP30	G	121	55	-	CTAAAGTCAAAAATGGCG
Saccharomyces							
M88698	YSCHSP150A	HSP150	G	?	?	?	AATAATAATATAATGCAA
M67479	YSCHSP104A	HSP104	G	?	?	?	TATACAGAATATATGAAC
K01387	YSCHSP90	HSP90	G	?	?	?	ACACGCAAAGATATGGCT
M26044	YSCHSP82	HSP82	G	?	?	?	TCATTCTGAAATATGGCT
M31006	YSCBIP	GRP78	G	39	59	-	ATCAAACATACCATGTTT
M98437	YSCHSP70X	HSP70	G	?	?	?	TTGCTGCTGCTAATGGGA
X01659	SCSSA1	HSP70	G	60	80	-	ATAACAGATAATATGTCA
X12927	SCSSA2	HSP70	G	54	76	-	AATATTTATACAATGTCT
M36115	YSCSSA3A	HSP70	G	25	68	-	CTACAAAGAAAAATGTCT
M32806	YSPHSP70	HSP70	G	104	53	3	CATGGAAGTCGCATGACC
J05637	YSCHSPSSA4	HSP70	G	156	71	1	AAAACAATAATCATGTCA
X59987	SPSSP1	HSP70	G	261	73	-	AACTCAAAAAATATGATC
M33301	YSCHSP60	HSP60	G	34	74	-	AAAGTTTTCAAAATGTTG
M93123	YSCHSP30X	HSP30	G	?	?	?	AACAAATTCAAATGAAC
M23071	YSCHSP26A	HSP26	G	?	?	?	AAACAAATTAACATGTCA
X55785	SCHSP	HSP12	G	60	75	-	CTAAATACAACAATGTCT
J03139	YSCHSF	HSTF	G	628	59	10	ATTGTTGGCGCCATGAAT
Histoplasma							
L11390	HTOHSP60X	HSP60	G	251	58	-	AATAGCTTCATCATGCAG
ALGAE							
Chlamydomonas							
M76725	CREHSP70A	HSP70	G	89	56	-	GAGTCAGTAAACATGGGC
X15053	CRHSP22K	HSP22	C	86	52	-	TCGCTAACCGACATGGCG
Pavlova							
X59555	PSTASP70	HSP70	G	?	?	?	AAGATAGAAAAAATGGCA

Explanation of the table headings:

a: GenBank/EMBL accession number; b: Locus; c: Heat shock induction related protein; d: Type of the sequence, C for cDNA and G for genomic; e: Length in bp; f: A + T%; g: Number of upstream ATGs; h: Context sequences of functional or proposed translation start codon.

Notes: HSP: Heat shock protein; HSC: Heat shock cognate protein; HSTF: Heat shock transcription factor; $ Genus name from the biological name has been mentioned and underlined.; #'G ? ? ?' indicates that genomic sequence has been isolated but no data has been given regarding the transcription start site or putative TATA box in the feature table of GenBank/EMBL data bank entries. $ and # are shown in parenthesis at their first occurrence.

Table 2. Data regarding 5' untranslated leaders of 10 representative non-HSP genes from four groups of eukaryotes.

a	b	c	d	e	f	g	h
Vertebrates							
Mus non-HSP genes							
D16497	Musbnp	Brain natri-uretic pro.	G	78	38	-	CGCTTCTGCGGCATGGAT
D16464	Mushes1	H-L-H	G	248	63	1	AAAAAAACGAAAATGCCA
L17322	Musent01	entactin	G	35	43	-	CAGTTGGGAAACATGCTG
X01153	Rnwap1	acidic Pro.	G	33	36	-	GCCGCCGACACCATGCGC
X03020	Mmgmcsfg	colony stim. factor	G	32	47	-	GTCCTGAGGAGGATGTGG
X66406	Mmgcol1	collagen	G	131	37	-	CCAGCCAGAAACATGAGG
X62302	Mmezg1	zeta globin	G	64	39	-	ACCACCACCACCATGTCT
X07295	Mmdh1	malate deh-drogenase	G	101	34	-	CCCGCCCTAGCCATGCTG
X56007	Mmatpb2	Na/K atpase	G	584	37	-	CGTGCCTCCAAGATGGTC
X07302	Mmaspat1	asp. amino transferase	G	87	34	-	CATTCTGTCGCGATGGCG
Invertebrates							
Drosophila non-HSP genes							
X12536	Dmadh12	Adh	G	68	56	1	AGCTCCATAGAAATGGCA
X04569	Dmamyag1	amylase	G	41	57	-	TCTGGAATCATCATGTTT
X13831	Dmsod	sod	G	115	66	-	CAAAACACAAAAATGGTG
U02879	Dmtfiib	Tfiib	G	978	61	21	TACAGAGTACAAATGGCA
X12506	Dmtwistg	twist	G	165	56	-	CAAGATCACCAAATGATG
U01335	Dmu01335	ribosomal protein	G	42	64	-	TAGACCTACAAAATGGCG
U07799	Dmu07799	Formalde-hyde dehy-drogenase	G	158	54	-	CAGACTGCAAAGATGTCT
X05427	Dmubxg5	ultra bithorax	G	20	43	-	CACATTCGTTCGATGGCA
M92914	Dromastm	mastermind	G	366	63	3	ACGGATGTATTTATGGAT
Higher plants							
Arabidopsis non-HSP genes							
L19637	Athaptla	APT	G	111	56	-	CGGGACAGTGAAATGGCG
L14749	Athg3p	GLPDH	G	46	60	-	TCACTTGCAGCTATGGCC
L33678	Athpal2	PAL2	G	112	64	-	TCTTGAAAACCAATGGAT
L19262	Athnadph	HMG red.	G	200	62	3	TCTCCGGCGTCAATGGAG
D01113	Athrd22	RD22	G	49	63	-	CAACTCCCAAAAATGGCG
U11033	Atul1033	Invertase	G	95	72	-	TCTCTTGCACCAATGAGT
L28828	Athmtripr	ribo.prot.	G	106	50	-	TCAACATCAGCCATGGCG
L24070	Athcor15b	cor15	G	84	65	1	GTCCTCATGGCGATGTCT
L34685	Athaptr	APTR-1	G	46	68	-	ACAACAAGAAAAATGAAG
M95594	Athacs	ACS	G	99	67	-	TTGACACAGCAAATGGGT
Lower plants							
yeast non-HSP genes							
U00145	U00145	catalase	G	142	61	-	ATTTTAAGCAAAATGGCT
D14538	Ysrsmk	smk1	G	218	63	1	AACTTACACAGCATGCGT
X05107	Spcdc17	DNA Ligase.	G	17	65	-	GACTTGATTGTTATGCGA
X07561	Scura4	DH orotase	G	41	63	-	CCTAACAAAGCAATGGTA
X03010	Scste2	phrom. Res.	G	32	63	-	CAAGAATCAAAAATGTCT
X12775	Scsrp1	Ser.Rich Pr.	G	61	75	-	CAATAATTAAAAATGGCT
X12576	Scmal6r	Malt. Ferm.	G	83	71	-	TCATATAATAATATGGGT
X03977	Scfas1	FA synth.	G	36	58	-	TATTCGCTCATTATGGAC
X06790	Sccargb	O.A transf.	G	56	52	-	ACCACAACACCAATGTCC
X05992	Scadh4	ADH	G	62	73	-	CAAATAAGAAAAATGTCT
X00430	Kllac4	LAC4	G	115	69	-	GAACTGAAAGATATGTCT

Explanation of the table headings:
a: GenBank/EMBL accession number; b: Locus; c: Abbreviated gene name; d: Type of the sequence, C for cDNA and G for genomic; e: Length in bp; f: A + T%; g: Number of upstream ATGs; h: Context sequences of functional or proposed translation start codon.

except that there are 5 bp between the end of the 5′ intron and beginning of the ATG. No plant HMW HSP genes share these characteristics with their animal counterparts.

Lengths of HSP leaders

While HSP leaders are alleged to be unusually long (11,29), our data do not support this assumption. The distribution patterns of HSP leader lengths of 139 sequences included in Table 1 (column e) show that majority (77%) of these sequences are <150 bp long, 20% sequences are in the range of 151–300 bp and only 3% are >300 bp. Diagrammatic representation of 5′ UTL lengths in different groups of eukaryotes for vertebrates (Fig. 1a), invertebrates (Fig. 1b), higher plants (Fig. 1c) and lower plants (Fig. 1d) indicate that different patterns in 5′ UTL lengths are present in these groups. While 26% of the vertebrate HSP 5′ UTLs are longer than 150 bp, 43% of the invertebrate 5′ UTLs are longer than 150 bp. In the case of higher plants, 92% of the 5′ UTLs are shorter than 150 bp and 27% of the representative lower plant genes have 5′ UTLs above 150 bp. These data are similar to earlier compilations of HSP as well as non-HSP genes from vertebrates (13) and plants (14) where common lengths of 5′ UTLs were below 200 nucleotides. Moreover, when a comparison of 5′ UTLs between about 10 or more HSP genes (Table 1, column e) and 10 non-HSP from mouse, drosophila, arabidopsis and yeast (Table 2, column e) was made for their lengths, both groups showed similar lengths (mostly <200 bp) with a few gene-specific exceptions. Deletion analysis using leader sequences from *Drosophila* HSP22 indicated that its first 22 nucleotides were most important for the preferential translation at high temperatures whereas the rest of the leader was dispensable (6,7). How much of the sequence of HSP mRNA leaders is really important for its function in preferential translation at high temperatures has not been extensively investigated. The length of HSP 5′ UTLs are similar to other cellular mRNAs.

Base composition of HSP leaders

GC richness of a 5′ UTL sequence is indicative of the stability of the potential secondary structures whose formation may hinder the progress of 43S initiation complex during the scanning mechanism of translation initiation (12). In fact, leader sequences of most of the vertebrate mRNAs are GC rich (50–70%) and therefore, more structured and undertranslated. On the other hand, plant and yeast mRNA leaders are AT or A rich and most of these mRNAs are unstructured and may not inhibit the forward movement of the initiation complex from the cap to functional ATG. On the basis of information from *Drosophila* HSP mRNAs, it was assumed that HSP leaders are AT rich and pose fewer complications regarding potential secondary structures. The present compilation clearly indicates that this is not the case. Figure 2a shows that leaders from vertebrate mRNAs are significantly GC rich (83% of the HSP mRNA leaders) whereas invertebrate mRNAs (Fig. 2b) have 94% HSP mRNAs that are AT rich. In the case of higher plants, only 19% HSP mRNA leaders are GC rich and the remaining 81% are AT rich (Fig. 2c). Interestingly, all the HSP mRNA leaders from lower plants are highly AT rich (Fig. 2d). These data do not deviate from previous observations with these members of eukaryotes indicating that HSP mRNA leaders do not have special base composition attributes and that they follow the same trends as other non-HSP mRNAs (13,14). These observations are further confirmed using a set of non-HSP genes as described in methods where vertebrate genes were mostly observed to be GC rich and invertebrate, and many plant genes were AT rich.

Secondary structural predictions could give some idea about the unstructured nature of 5′ UTLs of HSP mRNAs. Given the number of sequences in this compilation, it is not only difficult to perform this analysis with all these sequences but such computer simulations are also of questionable value. We have selected a few sequences from our compilation for these studies. 5′ UTLs with high GC content tend to show more structured character than AT rich sequences (data not shown).

Context sequences of the upstream and functional ATG codons

As discussed in the Introduction, sequences flanking the ATG codon are believed to be important for the efficiency of the translation initiation. This is the region which could show major

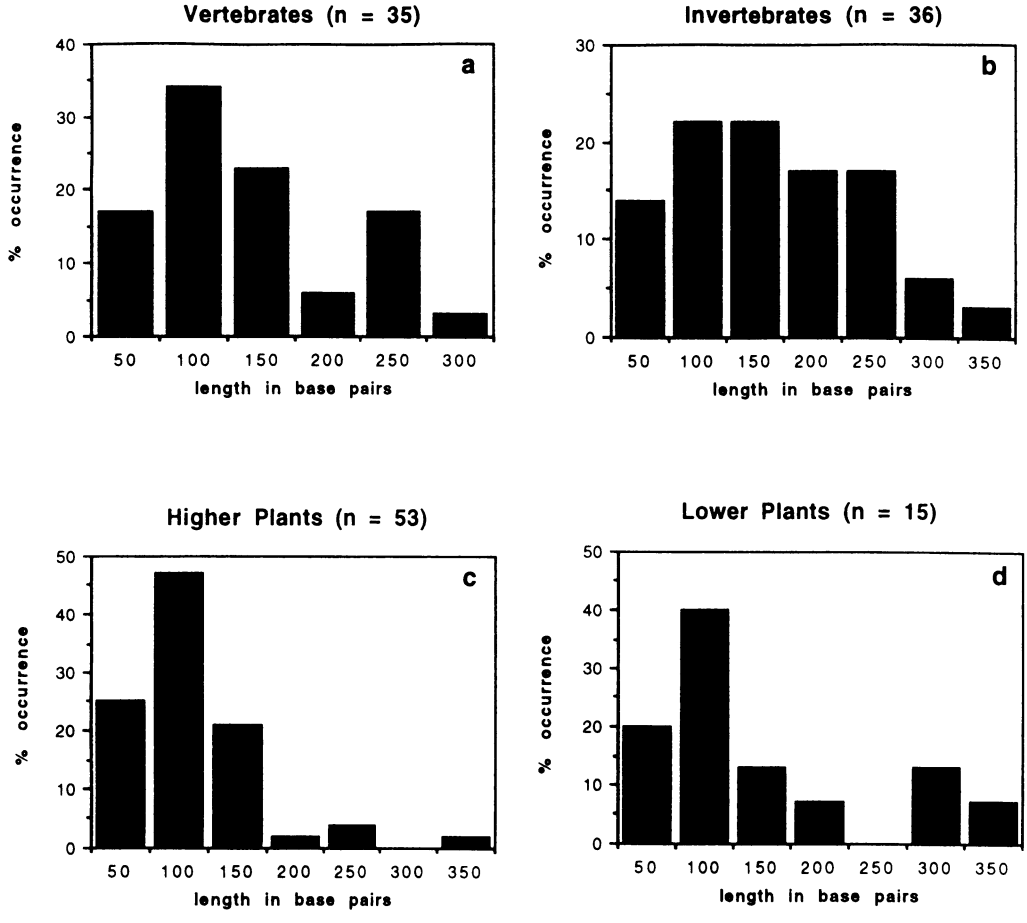

Figure 1 Length distribution of the 5′ untranslated leaders of HSP mRNAs from four groups of eukaryotes namely, (**a**) vertebrates, (**b**) invertebrates, (**c**) higher plants and (**d**) lower plants.

deviations from the normal trend and may be involved in the preferential translation of HSP mRNAs during heat stress. Table 3 shows the frequencies of base occurrence in the 12 bp upstream and 3 bp downstream regions of the proposed functional ATG codons from HSP gene among vertebrates, invertebrates, higher plants and lower plants. Table 4 presents the data regarding non-HSP genes from the same four groups. Interestingly, the consensus sequences of these regions are highly similar to the previous compilation using HSP as well as non-HSP mRNAs (13,14,17). The common occurrence of a G residue at –6, –9 and –12 positions in vertebrate mRNAs as discussed by Kozak (13) is also clearly evident. These data fail to reveal a trend in base arrangement and suggest that the context sequences of HSP 5′ UTLs are similar to other eukaryotic genes. These observations are reconfirmed by comparing the data between HSP and non-HSP genes from mouse/rat, *Drosophila*, *Arabidopsis* and yeast.

An attempt to deduce a consensus sequence for the context of functional ATG in the different gene families was also made. For example, ATGs of HSC70 genes from vertebrates are preceded by GCAACC, a feature not conserved in invertebrates or plants. Vertebrate GRP78 mRNAs have GGCAAG preceding the functional ATG whereas this feature is absent in other organisms. The context of translation initiation codon of HSP90 mRNAS from vertebrates is YCAAG and invertebrates have ATACAAG. These observations appear to have more evolutionary implications than functional significance.

5′ UTLs of polioviruses and other picorna viruses sharing the same behavior as HSP mRNA leaders are characterized by multiple upstream ATG codons. The data compiled here is the ideal place to determine the presence of upstream ATGs in HSP leaders. Only 19 HSP mRNAs (14%) of the 139 sequences listed in Table 1 (column g) have upstream ATGs. The majority of these ATGs (80%) are followed by in-frame stop codons before reaching the proposed functional ATGs and 57% of the upstream ATGs have A/G at –3 position which should have been selected for translation if the good context hypothesis of Kozak (29) is adequate. However, this observation is not unique to HSP mRNA leaders. Table 2, column g shows that occurrence of upstream ATGs is similarly common among non-HSP genes from different eukaryotic groups and is not a specialty of HSP genes. We have also attempted to search if separate classes of HSP leaders can be identified in this database using GCG program 'Pileup'. No clear classes of sequences could be identified. In several cases, highly similar sequence motifs of ≤ 10 bp occurred in sequences from different and unrelated organisms. However, we do not know the significance of such observations.

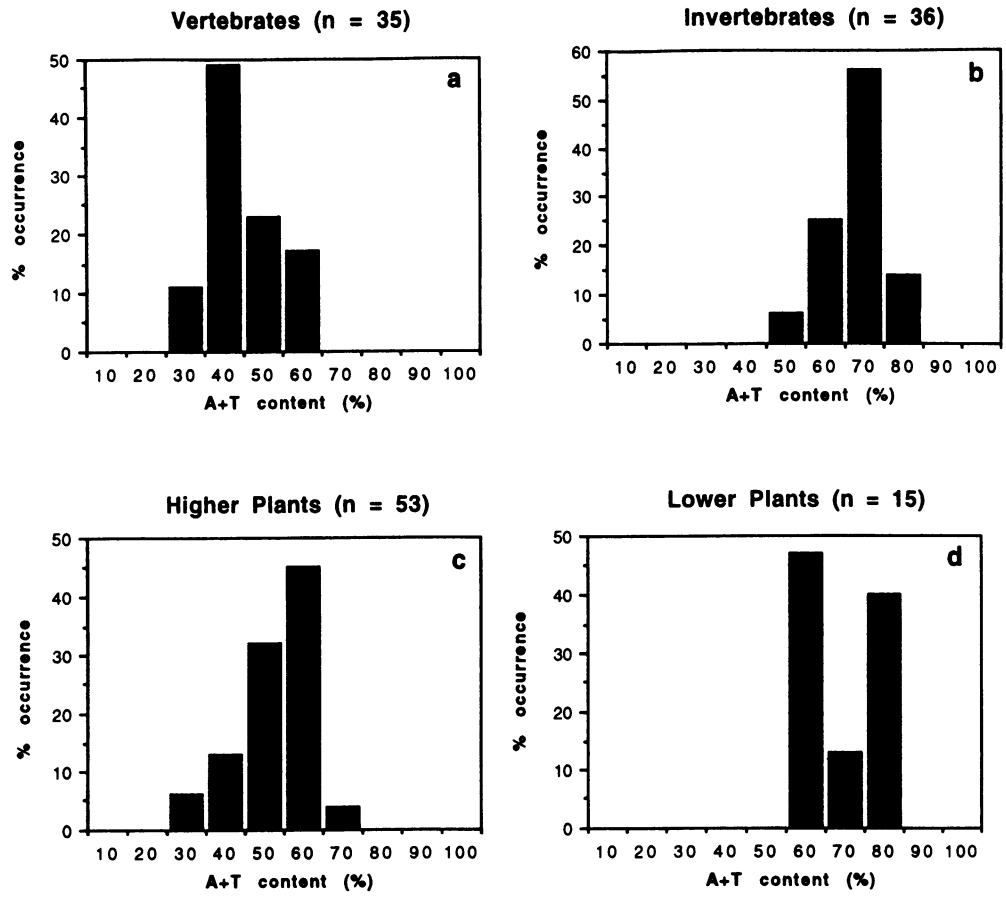

Figure 2 Base composition (A+T%) of the 5′ untranslated leaders of HSP mRNAs belonging to vertebrates (Fig. 1a), invertebrates (Fig. 1b), higher plants (Fig. 1c) and lower plants (Fig. 1d).

Sequence motifs complementary to conserved 3′ end of eukaryotic 18S rRNA

The 3′ end of 18S rRNA is highly conserved in eukaryotes and has been proposed to participate in the process of correct functional ATG selection (17,30). As discussed in the Introduction, the IRES in poliovirus mRNAs was also suggested to be the nucleotide sequence complementary to highly conserved 18S rRNAs (24). If HSP mRNA leaders show preferential translation during heat shock similar to polioviral mRNAs, then similar sequence motifs complementary to 18S rRNA ends could be speculated to be exclusively present in HSP mRNA leaders. A search was therefore made for a short sequence motif (5 bp) which is perfectly complementary to the sequence at the 3′ end of 18S rRNA (31). A 33 bp region from the conserved 3′ end was arbitrarily selected and 29 overlapping short sequence motifs were deduced. The GCG Program routine 'Find' was employed to search the occurrence of perfect matches with these motifs in 139 5′ UTLs from HSP mRNAs. A total of 107 (77%) sequences showed presence of one or more sequence motifs from the search profile of 29 motifs. The variety of motifs present in these sequences indicate that although HSP translation is a general phenomenon, the exact mechanism of internal entry of ribosome could be a species- and gene family-specific process. Moreover, there exists a great level of redundancy in the occurrence of sequence motifs if these sequences are responsible for internal initiation of translation. It could be speculated that after the entry of the 40S subunit on the 5′ UTL, normal linear scanning of the leader takes place as described by Kozak (29). The absence of motifs may be due to one of the following reasons: (i) cDNAs that are incomplete at the 5′ end or ill defined putative transcription start site of genomic sequences, (ii) perfectly matching short sequence motifs may not always be necessary or (iii) other mechanisms of translation initiation may exist in these mRNAs.

Due to the diverse nature of eukaryotic species and gene families reported in this study, a single mechanism of preferential HSP mRNA translation might be difficult to deduce. Interestingly, in a few cases a perfectly matching (complementary) motif of 6–8 and even up to 11 bp was found. This suggests a possibility that such mechanism may exist in nature. However, there is one major problem for this speculation. Most of the non-HSP genes included in this survey do not show absence of these 5 bp motifs and this adds into the complexity of the problem regarding why HSP genes are preferentially translated during heat stress. Perhaps these motifs become important as IRES only in the absence of mRNA cap and non-HSP mRNAs that are already capped are not subjected to IRES selection. Moreover, new synthesis of non-HSP transcripts is inhibited during heat stress. However, it must be stated that this needs to be confirmed experimentally.

The subject of preferential translation of HSP mRNAs from eukaryotes during heat stress has been extensively investigated

Table 3. Nucleotide frequencies at positions flanking the start codon of HSP genes

VERTEBRATES (n = 38)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	24	29	26	3	18	24	29	18	32	76	45	13	100	-	-	24	32	3
C%	29	32	24	24	58	32	8	37	47	5	34	42	-	-	-	29	58	24
G%	29	16	24	45	11	29	45	24	18	18	18	39	-	-	100	18	-	34
T%	18	24	26	29	13	16	18	21	3	-	3	5	-	100	-	29	11	39
*	c g	c	a t	g	C	c	g	c	C A	A	A C	C G	A	T	G	c g	C A	t
**	NA	NA	c	g	c	c	g	c	c	A G	c	C	A	T	G	g	c	NA

INVERTEBRATES (n = 44)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	43	52	34	68	34	57	25	30	41	82	57	45	100	-	-	20	11	32
C%	20	16	20	5	39	11	20	32	43	5	18	11	-	-	-	34	77	18
G%	20	14	18	14	7	-	25	9	5	11	16	30	-	-	100	18	7	18
T%	16	18	27	14	20	32	30	30	11	2	9	14	-	100	-	27	5	32
*	a	A	a	A	c	A T	t	c	C A	A	A	A G	A	T	G	c	C	a t
**	NA	NA	a	a	c	a	a	c	C A	A	a	a c	A	T	G	g	c	NA

HIGHER PLANTS (n = 59)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	27	25	44	27	27	34	20	32	53	54	36	51	100	-	-	5	15	10
C%	22	27	12	25	25	19	31	20	10	12	42	17	-	-	-	-	81	15
G%	24	12	10	20	20	19	29	19	22	29	7	24	-	-	100	56	3	25
T%	27	36	34	27	27	29	20	29	15	5	15	8	-	100	-	39	-	49
*	a t	t	A T	a t	a t	a	c	a	A	A G	C A	A	A	T	G	G T	C	t
**	NA	NA	a	a	a	a	a	g	a	A G	C A	a	A	T	G	G	C	NA

LOWER PLANTS (n = 25)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	60	56	36	48	28	52	40	40	44	72	52	44	100	-	-	20	20	24
C%	20	12	16	24	20	12	12	16	20	12	24	32	-	-	-	8	56	20
G%	4	4	16	12	16	8	20	16	8	12	4	-	-	-	100	40	12	16
T%	16	28	32	16	36	28	28	28	28	4	20	24	-	100	-	32	12	40
*	A	A T	a	a	t	A T	a	a	a	A	A	A C	A	T	G	g	C	t
**	NA	NA	a	a	a	a	a	a	a	A	a	a	A	T	G	t	c	NA

*Consensus sequence for HSP genes, determined by the criteria described by Cavener (16) and Joshi (27)

**Consensus sequence for Non-HSP and HSP genes determined on the basis of the data reported by Cavener and Ray (17). NA represents non-available data.

Table 4. Nucleotide frequencies at positions flanking start codon of non-HSP genes

VERTEBRATES (n = 10)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	20	30	20	20	10	10	40	10	30	70	40	10	100	-	-	10	10	10
C%	60	40	50	10	60	60	10	30	50	-	40	60	-	-	-	40	30	20
G%	20	20	10	40	-	20	30	40	20	30	20	30	-	-	100	30	30	50
T%	-	10	20	30	30	10	20	20	-	-	-	-	-	100	-	20	30	20
**	C	c	C	g	C	C	a	g	A C	A	A C	C	A	T	G	c	c g/t	G
*	c g	c	a t	g	C	c	g	c	C A	A	A C	C G	A	T	G	c g	C A	t

INVERTEBRATES (n = 10)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	30	60	20	50	30	30	40	20	50	50	70	60	100	-	-	10	10	30
C%	40	20	30	-	30	30	20	20	40	20	10	10	-	-	-	-	60	-
G%	-	10	30	30	20	-	30	20	-	10	-	20	-	-	100	70	-	30
T%	30	10	20	20	20	40	10	40	10	20	20	10	-	100	-	20	30	40
**	c	A	c g	A G	a c	t	a	t	C A	A	A	A	A	T	G	G	C T	t
*	a	A	a	A	c	A T	t	c	C A	A	A	A G	A	T	G	c	C	a t

HIGHER PLANTS (n = 10)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	10	10	40	30	10	40	40	20	60	20	40	70	100	-	-	20	30	-
C%	20	60	10	50	40	30	20	50	-	30	60	10	-	-	-	-	50	10
G%	10	10	20	10	10	10	30	20	30	40	-	10	-	-	100	70	20	50
T%	60	20	30	10	40	20	10	10	10	10	-	10	-	100	-	10	-	40
**	T	C	a	A C	C T	a	a	C	A G	g	C A	A	A	T	G	G	C A	G T
*	a t	t	A T	a t	a t	a	c	a	A	A G	C A	A	A	T	G	G T	C	t

LOWER PLANTS (n = 10)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	20	70	50	30	40	50	60	40	60	60	50	50	100	-	-	-	10	20
C%	40	30	30	10	20	10	20	20	20	10	20	10	-	-	-	20	50	20
G%	20	-	-	10	-	30	-	10	-	30	10	-	-	-	100	40	30	-
T%	20	-	20	60	40	10	20	30	20	-	20	40	-	100	-	40	10	60
**	c	A	A C	T A	A T	A G	A	a	A	A G	A	A T	A	T	G	G T	C G	T
*	A	A T	a	a	t	A T	a	a	a	A	A	A C	A	T	G	g	C	t

*Consensus sequence for HSP genes, determined in this paper (Table 3) by the criteria described by Cavener (16) and Joshi (17)

** Consensus sequence for Non-HSP genes determined on the basis of the data compiled and analysed in this paper.

for the last 12 years. Although we have made considerable progress and collected valuable information, the exact role of 5′ UTLs from HSP mRNAs in this process has remained elusive. In response to heat shock, a massive reprogramming of the transcription and translation processes takes place due to the changes in protein requirements. Prior to these changes, polysomes carrying normal mRNAs virtually disappear and most of the pre-existing cellular mRNAs stay in an unaltered state and intact conditions as evidenced by their translation *in vitro* or during the recovery periods (5). Why these mRNAs can not be recruited onto polysomes during heat stress is not clearly known although alterations in the translation initiation factors seems to be a strong possibility. During heat stress, transcription machinery gears up toward an almost exclusive synthesis of HSP mRNAs which is greatly assisted by the presence of the highly conserved heat shock element (HSE) in the promoter regions of heat shock protein genes and binding of heat shock transcription factor (HSTF) to HSEs. These newly formed mRNAs vigorously form new polysomes while normal mRNAs cannot be translated during heat stress. It is, therefore, possible that HSP mRNAs have a different mechanism of translation initiation. The HSP 5′ UTL seems to play a vital role in this process by allowing the internal and cap-independent entry of the 40S subunit. The unanswered question is what structural features of the primary HSP mRNA sequence are responsible for their escape from thermal repression. As discussed here, the length and GC content of 5′ UTLs per se may not be important for this behavior. Data presented in this paper do not support the possibility that 5′ UTLs have unique features which make them more successful than non-HSP mRNAs in this process. Although the presence of multiple sequence motifs complementary to the 3′ end of 18S rRNA is one of the possible explanations, clear experimental evidence is lacking. The small size of the motif could be one reason why defining the exact translation element has proved to be frustrating and more research in this direction could be rewarding (32). Unlike HSEs, translation elements appear to be less conserved among eukaryotes, however a high degree of species and gene family specificity exist (5,9). There is also a need to obtain more evidence concerning internal initiation of HSP mRNAs since several mechanisms might exist among different eukaryotes and different HSP gene families.

ACKNOWLEDGEMENTS

We are grateful to Dr Daniel Gallie, University of California, Riverside, for critical reading of this manuscript and our several colleagues at Texas Tech University for their helpful suggestions. This work was partly supported by a grant from US Department of Agriculture National Competitive Grants Program (92-37100-7903). Contribution of the College of Agricultural Sciences, Texas Tech University, Journal number T-4-380.

REFERENCES

1 Lindquist, S. and Craig, E.A. (1988) *Annu. Rev. Genet.* **22**, 631–677.
2 Vierling, E. (1991) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **42**, 579–620.
3 Nguyen, H.T. and Joshi, C. P. (1993) In Mabry, T.J., Nguyen, H.T., Dixon, R.A., and Bonness, M.S. (eds) *Biotechnology for Aridland plants*. IC^2 Institute, Austin, pp. 93–105 .
4 Welsh, W.J. (1993) *Sci. Am.*, **268** , 56–64.
5 Nover, L., Neumann, D., Scharf, K-D. (1989) *Heat Shock and other Stress Response Systems of Plants*. Springer-Verlag, Berlin.
6 Hultmark, D., Klemenz, R. and Gehring, W.J. (1986) *Cell* **44**, 429–438.
7 McGarry, T.J. and Lindquist, S. (1985) *Cell* **42**, 903–911.
8 Klemenz, R., Hultmark, D., and Gehring, W.J. (1985) *EMBO J.* **4**, 2053–2060.
9 Pitto, L., Gallie, D.R., and Walbot, V. (1992) *Plant Physiol.* **100**, 1827–1833.
10 Gallie, D.R. and Walbot, V. (1992) *Nucleic Acid Res.* **20**, 4631–4638.
11 Lindquist, S. (1986) *Annu. Rev. Biochem.* **55**, 1151–1191.
12 Kozak, M. (1992) *Annu. Rev. Cell Biol.* **8**, 197–225.
13 Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148.
14 Joshi, C.P. (1987) *Nucleic Acids Res.* **15**, 6643–6653.
15 Hamilton, R., Watanabe, C.K., and de Boer, H.A. (1987) *Nucleic Acids Res.* **15**, 3581–3593.
16 Cavener, D.R. (1987) *Nucleic Acids Res.* **15**, 1353–1361.
17 Cavener, D.R. and Ray, S.C. (1991) *Nucleic Acids Res.* **19**, 3185–3192.
18 Merrick, W.C. (1992) *Microb. Rev.* **56**, 291–315.
19 Thach, R. E. (1992) *Cell* **68**, 177–180.
20 Lamphear, B.J. and Panniers, R. (1991) *J. Biol. Chem.* **266**, 2789–2794.
21 Joshi-Barve, S., Bendetti, A.D., and Rhoads, R.E. (1992) *J. Biol. Chem.* **267**, 21038–21043.
22 Pelletier, J. and Sonneberg, N. (1988) *Nature*, London **334**, 320–325.
23 Jackson, R.J., Howell, M.T. and Kaminski, A. (1990) *Trends Biochem. Sci.* **15**, 477–483.
24 Pilipenko, E.V., Gmyl, A.P., Maslova, S. V., Svitkin, Y.V., Sinyakov A.N., and Agol V.I. (1992) *Cell* **68**, 119–131.
25 Macejak, D.G. and Sarnow, P. (1991) *Nature*, London **353**, 90–94.
26 Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
27 Joshi, C.P. (1987) *Nucleic Acids Res.* **15**, 9627–9640.
28 Devereux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
29 Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857–872.
30 Sargan, D.R., Gregory, S.P., and Butterworth, P.H.W. (1982) *FEBS Lett.* **17**, 133–136.
31 Dams, E., Hendriks, L., Van de Peer, Y., Neefs, J-M., Smits, G., Vandenbempt, I. and de Wachter, R. (1988) *Nucleic Acids Res.* **16**, r87.Vandenbempt, I. and de Wachter, R. (1988) *Nucleic Acids Res.* **16**, r87.
32 Yost, H.J., Petersen, R.B., and Lindquist, S. (1990) In Morimoto, R.J., Tissieres, A. and Georgopoulos, C. (eds) *Stress Proteins in Biology and Medicine*, Cold Spring Harbor Press. pp. 379–409.