J Mol Evol (1994) 39:80–86

Journal of Molecular Evolution

Phylogenetic Analysis of the Stress-70 Protein Family

S.A. Rensing, U.-G. Maier

Institut für Biologie II, Schänzlestr. 1, D-79104 Freiburg, Germany

Received: 17 May 1993 / Revised: 20 November 1993 / Accepted: 5 December 1993

Abstract. The eukaryotic cyto-/nucleoplasmatic 70-kDa heat-shock protein (HSP70) has homologues in the endoplasmic reticulum as well as in bacteria, mitochondria, and plastids. We selected a representative subset from the large number of sequenced stress-70 family members which covers all known branches of the protein family and calculated and manually improved an alignment. Here we present the consensus sequence of the aligned proteins and putative nuclear localization signals (NLS) in the eukaryotic HSP70 homologues. The phylogenetic relationships of the stress-70 group family members were estimated by use of different computation methods. We present a phylogenetic tree containing all known stress-70 subfamilies and demonstrate the usefulness of stress-70 protein sequences for the estimation of intertaxonic phylogeny.

Key words: Heat-shock proteins — HSP70 — Stress-70 — Phylogenetic trees — Protein families — NLS

Introduction

When the first heat-inducible chaperones were discovered in eukaryotes they were termed heat-shock proteins. Those belonging to the 70-kDa group (ranging from about 68 to 78 kDa) were called 70-kDa-class heat-shock proteins (HSP70). Today it is known that most of the chaperones which are expressed under heat shock or other stress conditions have a counterpart which is expressed constitutively or under nonstress conditions. These proteins are termed heat-shock cognates. Besides these members the HSP70 family also contains proteins which do not have a known heat-inducible counterpart—for example, the glucose-regulated protein of the 78-kDa class (GRP78) which resides in the endoplasmic reticulum (ER) and which was identified initially as the immunoglobulin heavy-chain binding protein (BiP). The prokaryotic homologues of HSP70s are the DnaK chaperones, which were discovered because of their involvement in phage λ and *E. coli* DNA replication. (See, e.g., Friedman et al. 1984.) There are also close relatives of DnaK in plastids and mitochondria.

Unfortunately the terms are multifarious and some of them are ambiguous or describe parts of the family only. In addition, many of the terms are sometimes used in different ways. To unravel this confusion of terms all the chaperones of about 70 kDa have been recently termed stress-70 proteins (Gething and Sambrook 1992).

The functions of stress-70 proteins are multiple (Gething and Sambrook 1992; Georgopoulos 1992), ranging from binding to nascent polypeptide chains to disaggregation of protein agglutinates. In spite of their different functions and widespread distribution among all organisms, it is easy to identify the family members because of the high degree of conservation of these molecules. On the amino-acid level even distant relatives such as *E. coli* DnaK and human HSP70 share 46% identical positions and are 65% similar. This high degree of conservation, furthermore, allows the unambiguous alignment of homologous positions of the sequences.

We are concerned with understanding the molecular evolution of species as well as of their proteins. As we have recently cloned and sequenced the gene of a stress-70 protein from the nucleomorph (the vestigial nucleus

Correspondence to: S.A. Rensing

of the eukaryotic endosymbiont) of *Pyrenomonas salina* (Hofmann et al. 1994), we decided to determine the phylogenetic relationships in this protein family.

Methods

The search for sequences of stress-70 family members was carried out with the IRX semantic retrieval system at HUSAR (Heidelberg Unix Sequence Analysis Resources). All of the EMBL, GenBank, SWISS-PROT, and PIR databases (newest releases available in March 1993) were scanned for respective entries. A total of 90 sequences was found within this search. A subset of 33 sequences was selected which—to our best knowledge—cover all subfamilies known so far. No partial sequences were used. A portion of the Beet Yellows Closterovirus genome with significant similarity to *hsp*70 was chosen as outgroup. This sequence perfectly fits into the criteria for an outgroup, as it definitively is homologous to the sequences analyzed but shares less identical positions with every member of the family than any of the family members with each other. Table 1 lists the sequences used with the respective references and additional information.

The alignment was performed using CLUSTAL V (Higgins et al. 1992) and was manually improved by visual inspection afterward. The targeting signals of the nuclear-encoded organellar proteins were removed according to the information in the original literature and by cleavage analysis (performed with PC/GENE release 6.60; IntelliGenetics, Inc.). Two regions at the N- and C-terminal end of the sequences, of about 14 and 120 positions, which, due to large differences, could not be aligned unambiguously, were removed from the alignment.

A consensus sequence of the remaining 580 positions (insertions due to one sequence only were also deleted) has been calculated by use of the program CONSENSE (Rensing, available at the EMBL file server or on request).

The search for nuclear localization signals was carried out using PROSITE (rel. 10.0 of December 1992, Bairoch 1992) and by visual inspection.

Phylogenetic analysis was done with the program package PHYLIP 3.5c (unpublished; see Felsenstein 1989 for version 3.2); 597 positions of the alignment (without the ambiguous regions mentioned above) were used for the calculation. The estimation was carried out by Felsenstein's protein parsimony method, which does not count synonymous changes in one case and calculation of a protein distance matrix with neighbor-joining (Saitou and Nei 1987) afterward in the other case. Bootstrap resampling (Felsenstein 1985; Wu 1986) of the original data set was used as a pseudo-empirical test of the stability of the tree topology in either case. The consensus trees were constructed by use of the majority-rule and strict consensus algorithm implanted in PHYLIP.

Results

Searching the EMBL, GenBank, PIR and SWISS-PROT sequence databases with the information-retrieval system IRX, we found 90 *hsp*70 sequences, of which some are only partially available. We selected 33 sequences which to our knowledge cover all subfamilies of the multigene family known so far (Table 1).

As the amino-acid sequences contain the structural information of the proteins and are less "noisy" than the nucleic acid sequences, we chose them for our studies. In order to align the homologous regions of the proteins properly, we initially carried out a multiple sequence alignment with CLUSTAL V (Higgins et al. 1992) using the standard parameters and the PAM 250 matrix (Schwartz and Dayhoff 1978). The alignment was afterward improved in some points by visual inspection. The respective targeting sequences of the nuclear-encoded organellar proteins were deleted. For the GRP78 group sequences from human, yeast, and rat, the amount of discarded amino acids was 18, 42, and 18, respectively. In the case of the mitochondrial DnaK homologues from *Trypanosoma cruzi,* pea, and yeast 20, 44, and 23 amino acids were deleted, respectively, and in the case of the plastidal DnaK homologue from pea, 65 amino acids were deleted. We found that the very first part of the alignment (14 positions, about four to 13 amino acids) and a larger C-terminal region (120 positions, about 60–120 amino acids) were weakly conserved, and therefore it was not possible to align the sequences unambiguously in these regions. These two regions were taken into account neither for the consensus sequence nor for the phylogenetic trees, leading to a total length for the alignment of 597 positions.

A data set of 580 positions where the insertions due to only one sequence were also discarded was used for the calculation of the consensus sequence (Fig. 1). The respective position has to be conserved in at least 17 of the 33 sequences (52%) to be displayed in the consensus sequence. The division of amino acids into groups was carried out according to Ausubel et al. (1989); see figure legend for details. The consensus sequence revealed that there is one weakly conserved region (borders marked by arrows in Fig. 1) with no conservation between the subfamilies. Nevertheless, there is quite a high level of conservation in this area in the single subfamilies. The identified protein domains (see, e.g., Milarski and Morimoto 1989; Gething and Sambrook 1992) for ATPase activity and protein binding are not found to be reflected in the consensus sequence. The HSP70 family signatures according to Bairoch (PROSITE 10.0 December 1992, Bairoch 1992) can be found in the consensus sequence (marked as underlined in Fig. 1) but the family signature I needs to be expanded from [IV]-D-L-G-T-T-x-S to [IVL]-D-[LF]-G-T-T-x-S to fit all the sequences and an additional leading G may be added, whereas signature II remains D-[LF]-G(3)-T-F-D. We define two new universal stress-70 signatures of seven and six amino acids in length with the profile [TS]-[VC]-P-A-[YN]-[FY]-N and [NP]-[EG]-P-[TS]-A-A and, further, a signature specific to eukaryotic stress-70 proteins (HSP70, HSC70, and GRP78) with a length of seven amino acids and the profile R-A-[RK]-F-E-[ED]-[LM], which is located in the subfamily-specific region mentioned above. (All three double underlined in Fig. 1).

As the nucleo-/cytoplasmatic stress-70 proteins have to be transported to the nucleus we carried out a search with PROSITE for the bipartite motif proposed by Dingwall and Laskey (1991) to serve as nuclear localization

Table 1. The stress-70 sequences used for the alignment are listed[a]

Organism	Type	Cog.	Appr. MW	Reference	Acc. No.
Eukaryotic stress-70 group					
Yeast					
Yeast	SSA1		70 kDa	Slater and Craig (1989a)	X12926
Yeast	SSA2		70 kDa	Slater and Craig (1989a)	X12927
Yeast	SSA4		70 kDa	Boorstein and Craig (1990)	J05637
Yeast	SSB1	C	67 kDa	Slater and Craig (1989b)	X13713
Vertebrates					
Human	HSP70		70 kDa	Hunt and Morimoto (1985)	M11236
Human	HSP70B		71 kDa	Leung et al. (1990)	X51757
Human	HSC70	C	71 kDa	Dworniczak and Mirault (1987)	Y00371
Mouse	HSC70	C	71 kDa	Giebel et al. (1988)	M19141
Xenopus laevis	HSP70		71 kDa	Bienz (1984)	A03310 (PIR)
Drosophila melanogaster	HSC70	C	71 kDa	Perkins et al. (1990)	M36114
Plants					
Maize	HSP70		71 kDa	Rochester et al. (1986)	X03658
Soybean	HSP70		71 kDa	Roberts and Key (1991)	S14992 (PIR)
Petunia	HSP70		71 kDa	Winter et al. (1988)	X06932
Algae					
Chlamydomonas reinhardtii	HSP70		71 kDa	Müller et al. (1992)	M76725
Pyrenomonas salina	HSP70		72 kDa	Hofmann et al. (1994)	X72621
Protozoans					
Leishmania donovani	HSP70		71 kDa	MacFarlane et al. (1990)	X52314
Trypanosoma cruzi	HSP70		74 kDa	Engman et al. (1989)	X07083
GRP78/KAR group					
Yeast	KAR2	C	74 kDa	Nicholson et al. (1990)	M25394
Rat	GRP78	C	72 kDa	Munro and Pelham (1986)	M14866
Human	GRP78	C	72 kDa	Ting and Lee (1988)	M19645
DnaK group					
Eubacteria					
Escherichia coli	DnaK		69 kDa	Bardwell and Craig (1984)	K01298
Bacillus subtilis	DnaK1		65 kDa	Wetzstein et al. (1990)	X16393
Chlamydia pneumoniae	DnaK		71 kDa	Kornak et al. (1991)	X60083
Caulobacter crescentus	DnaK		68 kDa	Gomes et al. (1990)	M55224
Cyanobacteria					
Synechocystis sp. PCC6803	DnaK		68 kDa	Chitnis and Nelson (1991)	M57518
Archaea					
Methanosarcina mazei	DnaK		66 kDa	Macario et al. (1991)	X60265
Plastids					
Pavlova lutherii	ptDnaK		69 kDa	Scaramuzzi et al. (1992)	X59555
Cryptomonas phi	ptDnaK		68 kDa	Wang and Liu (1991)	M76547
Pea	ptDnaK		76 kDa	Marshall and Keegstra (1992)	L03299
Mitochondria					
Yeast	SSC1		71 kDa	Craig et al. (1989)	M27229
Trypanosoma cruzi	mtDnaK		71 kDa	Engman et al. (1992)	M73627
Pea	mtDnaK		72 kDa	Watts et al. (1992)	X54739
Outgroup					
Beet yellow Closterovirus	HSP70		65 kDa	Agranovsky et al. (1991)	X53462

[a] In the row "Cog." constitutively expressed/cognate proteins are marked by a C. The row "Appr. MW" shows the theoretical molecular weight of the respective protein. The accession numbers are EMBL nucleic acid accession numbers unless cited differently.

signal (NLS). This motif consists of two positively charged amino acids spaced from a block of another five amino acids, of which at least three are positively charged by approximately 10 amino acids. The search revealed the localization of one NLS each, which fits the consensus in most of the eukaryotic HSP70 homologues at the same position (about 250 amino acids from the N-terminus). Figure 2 shows the part of the alignment where the NLSs are located. In the human HSP70B the NLS is slightly shifted to the C-terminus. In the HSP70 proteins of *Leishmania donovani* and *Chlamydomonas reinhardtii* no NLS could be detected within the search. However, visual inspection shows that the respective positions of the proteins would match a slightly modified consensus. (See Fig. 2 for details.) In the GRP78 and DnaK homologues no NLS could be detected, as was the case for the viral sequence.

Preliminary evolutionary trees were constructed by using Felsenstein's protein parsimony method, which has proven to be effective for analyses of intertaxonic (e.g., Maerz et al. 1992) as well as protein family (e.g., Griess et al. 1993) phylogeny before and by use of

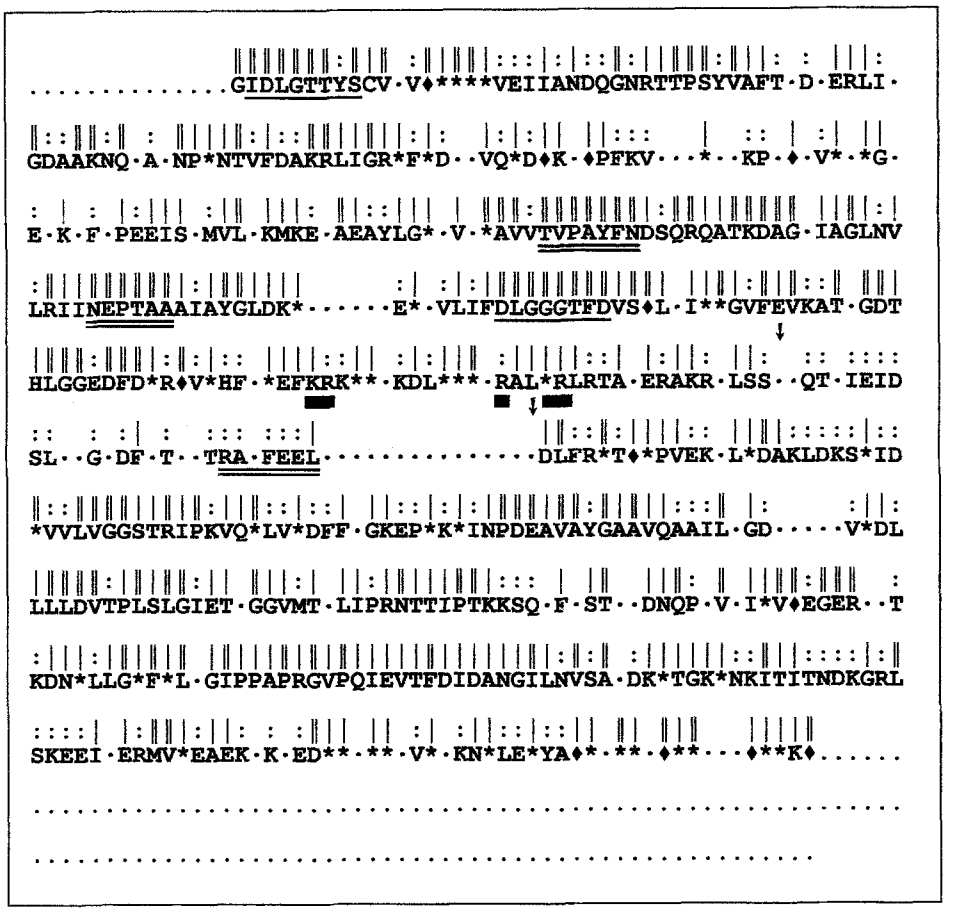

Fig. 1. Consensus sequence of the stress-70 group proteins. As mentioned in Methods, gaps due to one sequence only were removed before calculation, as were the ambiguous N- and C-terminal regions of the alignment (shown by *dots*). All 33 sequences were used to calculate the consensus sequence. The respective first lines show the conservation grade and the second lines the actual consensus sequence. The conservation grade lines contain symbols for the values 52% (:; 17 sequences), 76% (|; 25 sequences) and 97% (||; 32 sequences) and reflect the conservation of the amino acid or group shown below. The consensus sequence contains the amino acids in *one-letter code* if they are conserved in at least 17 sequences (52%). If conserved in at least 25 sequences (76%) symbols for hydrophobic (◆; containing amino acids C, F, Y, I, L, M, W, and V) and hydrophilic (⋆; containing amino acids S, R, H, G, K, Q, N, D, and E) amino acids are shown instead. Amino acids T, A, and P are considered neither hydrophilic nor hydrophobic. Positions which are conserved below 52% are shown by *dots* (·). The *arrows* mark the borders of a weakly conserved region in the middle of the sequences. The NLS consensus sequence (see Fig. 2 also) is shown by *bars*

neighbor-joining, as this method generally has a good chance of obtaining the correct evolutionary tree (Felsenstein 1988; Saitou and Imanishi 1989). The calculation of the protein distance matrix was carried out by three different methods: one according to Kimura (1980), one based on the PAM matrix of Schwartz and Dayhoff (1978), and one including a categories model of Felsenstein (PHYLIP 3.5, unpublished). Inspection of the four phylogenetic trees revealed that the protein parsimony tree was virtually identical with the Kimura neighbor-joining tree and that the PAM and categories neighbor-joining tree were identical, too. Differences between these two pairs of trees existed mainly in the grouping of the yeast SSB1, which branched off between GRP78 and HSP70 homologues in the parsimony/Kimura tree and between eu- and prokaryotes in the other two trees. Bootstrap analyses on the protein parsimony and the PAM matrix neighbor-joining tree were done to clarify the branching order. The bootstrap analyses lead to higher values for the SSB1 branching in the parsimony tree, which is presented in Fig. 3. The phylogenetic tree clearly shows a division in prokaryotic (DnaK and homologues) and eukaryotic (GRP78 and HSP70 homologues) sequences. The tree is largely in agreement with the study of nonlinear evolutionary rates in stress-70 genes carried out by Hughes (1993) with minimum evolution estimation (Rzhetsky and Nei 1992). In the eukaryotic part of the tree the GRP78/KAR2 sequences build a cluster which branches off first, immediately followed by the SSB1 sequence from yeast. The remaining eukaryotic sequences are divided into three main clusters: chlorobionts (maize to *Pyrenomonas*) and protozoans (*Leishmania* and *Trypanosoma*), animal "cognate" (human to *Drosophila*) and heat-shock proteins (human to *Xenopus*), and the fungal (yeast) sequences (SSA1, 2, and 4). The prokaryotic part of the tree is divided into two clusters, one consisting of mitochondrial DnaK homologues together

```
SSA4    Yeast   KRKNK-KDLTTNQRSLRR
SSA2    Yeast   KRKNK-KDLSTNQRALRR
SSA1    Yeast   KRKNK-KDLSTNQRALRR
HSC70   Mouse   KRKHK-KDISENKRAVRR
HSC70   Human   KRKHK-KDISENKRAVRR
HSC70   DROME   KRKHK-KDLTTNKRALRR
HSP70B  Human   RRKHG-KDLSGNKRALGRLRTACERAKRT
HSP70   Human   KRKHK-KDISQNKRAVRR
HSP70   XENLA   KRKHK-KDIGQNKRALRR
HSP70   TRYCR   KRKNKGKDLSTNLRALRR
HSP70   LEIDO   KRKNKGKNLASSHRALRG *
HSP70   CHLRE   QRKYK-KDLKTSPRALRR *
HSP70   GLYMA   KRKNK-KDISGNARALRR
HSP70   PETHY   KRKNK-KDISGNPRALRR
HSP70   Maize   KRKNK-KDISGNPRALRR
HSP70   PYSAL   KRKYK-KDVTSNARSLRR
HSC75   Yeast   KKKT-GLDISDDARALRR
GRP78   Human   KKKT-GKDVRKDNRAVQK
KAR2    Yeast   KKKH-GIDVSDNNKALAK
GRP78   Rat     KKKT-GKDVRKDNRAVQK
DnaK1   BACSU   KK-ENGVDLSKDKMALQR
DnaK    METMA   KK-SEGIDLSKDKAVLQR
DnaK    ECOLI   KK-DQGIDLRNDPLAMQR
DnaK    CHLPN   KK-QEGIDLSKDNMALQR
DnaK    CAUCR   KK-EQGVDLRKDKLALQR
DnaK    SYNSP   QK-AEGIDLRKDKQALQR
ptDnaK  PALUT   EK-EEKFSLKGDSQALQR
ptDnaK  CRYPH   ET-EHSINLKSDRQALQR
ptDnaK  Pea     KR-DEGIDLLKDKQALQR
mtDnaK  TRYCR   KK-STGIDLSNERMALQR
mtDnaK  Pea     KR-TESIDLAKDKLALQR
mtDnaK  Yeast   KT-ETGIDLENDRMAIQR
HSP70   BYV     NKAQLPVNYKIDISFL-K
                |::|| :|:|||| :||||
                KRK**-KDL***-RAL*R
```

Fig. 2. Putative nuclear localization signals (NLS) of the eukaryotic HSP70 homologues according to Dingwall and Laskey (1991). The region of the alignment containing the respective NLSs is shown, including the alignment consensus sequence. The consensus to which 13 of the 17 NLS-containing sequences fit is shown by *bars* under the consensus sequence as in Fig. 1. Positively charged residues which are part of the NLSs are shown in *bold*. The two sequences which do not exactly match the consensus are marked by *stars* (⋆). One has to suppose a histidine (H) as basic and nine amino acids as sufficient to space the basic domains from each other in order to fit the sequences with the proposed NLS consensus sequence respectively

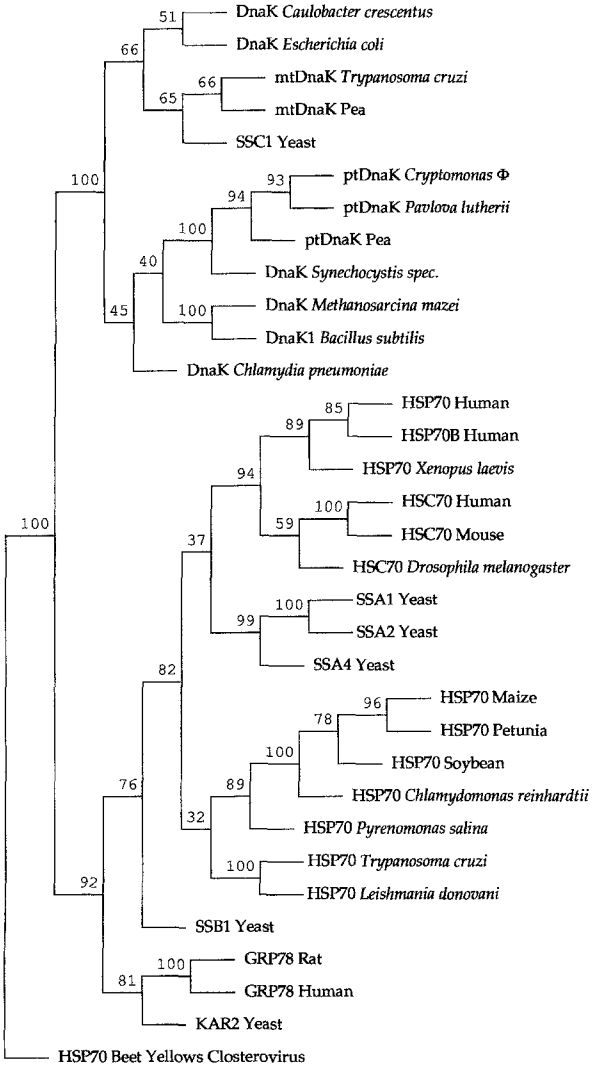

Fig. 3. Consensus phylogenetic tree of the stress-70 protein family. The tree is rooted by use of the outgroup. Bootstrap resampling and protein parsimony were carried out as mentioned in Methods. The phenogram contains the values from 100 times of bootstrapping. The respective values at the clusters show how often the group to the right was found. An average of 1.4 equal parsimonious trees was found throughout the bootstrap analysis. The number of required changes ranges from 4,285 to 4,819, with an average value of 4,591

with the DnaK sequences from *E. coli* and *Caulobacter crescentus* and the other of the plastidal DnaK homologues together with the DnaK of *Synechocystis, Methanosarcina mazei, Bacillus subtilis,* and *Chlamydia pneumoniae.*

Discussion

The analysis of the stress-70 protein sequences known today revealed that about 75% of the respective sequence length has been highly conserved between different lineages throughout evolution, reflecting the biological importance of the protein. Due to this high degree of conservation an unambiguous alignment is possible which enables the identification of far relatives from all kingdoms of organisms.

We have figured out three new stress-70 family signatures of which two ([TS]-[VC]-P-A-[YN]-[FY]-N and [NP]-[EG]-P-[TS]-A-A) are able to identify stress-70 proteins in general and one (R-A-[RK]-F-E-[ED]-[LM]) may serve as a signature to eukaryotic, nonorganellar stress-70 proteins. A SWISSPROT (rel. 24.0 from December 1992) database search was carried out with the mentioned motifs to prove their efficiency. Only motifs which were found in the nonpartial sequences (66 all in all) were taken into account. The search revealed that the motifs find the respective proteins to more than 95% (92% in the case of the eukaryote-specific signature); however, the signature [NP]-[EG]-P-[TS]-A-A matches a total of five wrong sequences, too. The modified signature I (G-[IVL]-D-[LF]-G-T-T-x-S) also has an efficiency of above 95% in determining stress-70 proteins.

The discovery of possible nuclear localization signals in the eukaryotic HSP70 homologues matching the proposed consensus is of great interest. The fact that the NLS motif is found in the HSP70 homologues only increases the probability that these patterns serve as the signal for nuclear targeting. If this is the case, one has to wonder if there can be a clearly separated function for "cognate" and stress-controlled proteins, as has been supposed by Craig and Jacobsen (1985) and others. (See Morimoto 1993.)

The evolutionary tree clearly shows the separation into two large divisions, one containing the eukaryotic nucleo-/cytoplasmatic and ER proteins and the other the prokaryotic DnaK proteins together with the plastidal and mitochondrial sequences. This branching order points out that an ancestral stress-70 gene must have existed before the separation into different kingdoms took place. The grouping of plastidal and mitochondrial proteins together with bacterial sequences is another proof for the endosymbiont theory (compare, e.g., Gray 1992). The nuclear-encoded genes of mitochondrial and plastidal stress-70 proteins (e.g., pea) must have been transferred into the nuclear genome some time ago, as there is no evidence for stress-70 genes in the mitochondrial and plastidal genomes of land plants and animals. The occurrence of these genes in the plastome of *Pavlova lutherii* (chromophytes) and *Cryptomonas* Φ (cryptophytes) fits in with the known "ancestry" of these genomes (see, e.g., Gray 1992) in comparison with, e.g., the plastids of the chlorobionts. The clustering of the different bacterial sequences with either plastidal or mitochondrial sequences is, at least in part, unexpected. It is not surprising that the cyanobacterial sequence clusters with the plastidal sequences, but the scattered order of the other bacterial sequences is not easy to understand. A possible explanation for these groupings, which do not seem to be in accordance with known systematics, is that there have possibly been (at least) two ancestral genes, of which one got lost in different lineages.

The position of the nucleomorph-encoded HSP70 from *Pyrenomonas salina* is at the base of the chlorobiont cluster. This is more relative to the protozoans than to higher plants, which seems to reflect a relative ancestry of the gene. This datum is in perfect agreement with former studies on the eukaryotic endosymbiont of cryptomonads (Maier 1992).

The branching of the yeast SSB1 protein as an outgroup for the HSP70 homologues is remarkable. Perhaps this position is due to a different function of this protein. (See Morimoto 1993.) The different gene control, organization, and possibly function of the stress-controlled and constitutively expressed proteins are reflected in the tree in the case of the animal HSP/HSC sequences. However, the differences between the two types of proteins are smaller than the respective differences between, for example, plant and fungal sequences, thus allowing the examination of intertaxonic relationships by use of these sequences.

Acknowledgments. We would like to thank Prof. P. Sitte for helpful discussions, Martin Walter (University computing center, Freiburg, FRG) for the kind permission to use workstation facilities, and Julia Hardiman for carefully reading the manuscript. This work has been supported by the Deutsche Forschungsgemeinschaft, grants to U.-G. Maier.

The program CONSENSE as well as the alignment are available on request by e-mail (INTERNET): rensing@sun1.ruf.uni-freiburg.de. The program package PHYLIP 3.5c is available by anonymous ftp from evolution.genetics.washington.edu.

References

Agranovsky AA, Boyko VP, Karasev AV, Koonin EV, Dolja VV (1991) Putative 65kDa protein of Beet Yellows Closterovirus is a homologue of HSP70 heat shock proteins. J Mol Biol 217:603–610

Ausubel FM (ed) (1989) Current protocols in molecular biology. J Wiley & Sons, Brooklyn, appendix 1, supplement 12

Bairoch A (1992) PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 20:2013–2018

Bardwell JCA, Craig EA (1984) Major heat shock gene of *Drosophila* and the *Escherichia coli* heat-inducible *dnaK* gene are homologous. Proc Natl Acad Sci USA 81:848–852

Bienz M (1984) *Xenopus hsp*70 genes are constitutively expressed in injected oocytes. EMBO J 3:2477–2483

Boorstein WR, Craig EA (1990) Structure and regulation of the SSA4 HSP70 gene of *Saccharomyces cerevisiae.* J Biol Chem 265: 18912–18921

Chitnis PR, Nelson N (1991) Molecular cloning of the genes encoding two chaperone proteins of the cyanobacterium *Synechocystis sp.* PCC6803. J Biol Chem 266:58–65

Craig EA, Jacobsen K (1985) Mutations in cognate genes of *Saccharomyces cerevisiae* hsp70 result in reduced growth rate at low temperatures. Mol Cell Biol 5:3517–3524

Craig EA (1989) *SSC1,* an essential member of the yeast HSP70 multigene family, encodes a mitochondrial protein. Mol Cell Biol 5:3517–3524

Craig EA, Kramer J, Shilling J, Werner-Washburne M, Holmes S, Kosic-Smithers J, Nicolet JM (1989) SSC1, an essential member of the yeast HSP70 multigene family, encodes a mitochondrial protein. Mol Cell Biol 9:3000–3008

Dingwall C, Laskey RA (1991) Nuclear targeting sequences—a consensus? Trends Biochem Sci 16:478–481

Dworniczak B, Mirault ME (1987) Structure and expression of a human gene coding for a 71kDa heat shock 'cognate' protein. Nucleic Acids Res 15:5181–5197

Engman DM, Sias SR, Gabe JD, Donelson JE, Dragon EA (1989) Comparison of *hsp*70 genes from two strains of *Trypanosoma cruzi.* Mol Biochem Parasitol 37:285–288

Engman DM, Fehr SC, Donelson JE (1992) Specific functional domains of mitochondrial HSP70s suggested by sequence comparison of the trypanosome and yeast proteins. Mol Biochem Parasitol 51:153–155

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521–565

Felsenstein J (1989) PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166

Friedman DI, Olson ER, Georgopoulos C, Tilly K, Herskowitz I, Banuett F (1984) Interactions of bacteriophage and host macromolecules in the growth of bacteriophage λ. Microbiol Rev 48: 299–325

Georgopoulos C (1992) The emergence of the chaperone machines. Trends Biochem Sci 17:295–299

Gething M-J, Sambrook J (1992) Protein folding in the cell. Nature 355:33–45

Giebel LB, Dworniczak BP, Bautz EK (1988) Developmental regulation of a constitutively expressed mouse mRNA encoding a 72-kDa heat shock-like protein. Dev Biol 125:200–207

Gomes SL, Gober JW, Shapiro L (1990) Expression of the *Caulobacter* heat shock gene *dnaK* is developmentally controlled during growth at normal temperatures. J Bacteriol 172:3051–3059

Gray MW (1992) The endosymbiont hypothesis revisited. Int Rev Cytol 141:233–357

Griess EA, Rensing SA, Grasser KD, Maier U-G, Feix G (1993) Phylogenetic relationships of HMG box DNA-binding domains. J Mol Evol 37:204–210

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved software for multiple sequence alignment. CABIOS 8:189–191

Hofmann CJB, Rensing SA, Häuber MM, Martin WF, Müller SB, Couch J, McFadden GI, Igloi GL, Maier U-G (1994) Smallest known eukaryotic genomes encode a protein gene: towards an understanding of nucleomorph functions. Mol Gen Genet (in press)

Hughes AL (1993) Nonlinear relationships among evolutionary rates identify regions of functional divergence in Heat-Shock Protein 70 genes. Mol Biol Evol 10:243–255

Hunt C, Morimoto RI (1985) Conserved features of eukaryotic *hsp*70 genes revealed by comparison with the nucleotide sequence of human *hsp*70. Proc Natl Acad Sci USA 82:6455–6459

Kimura M (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kornak JM, Kuo CC, Campbell LA (1991) Sequence analysis of the gene encoding the *Chlamydia pneumoniae* DnaK protein homolog. Infect Immun 59:721–725

Leung KC, Rajendran MY, Monfries C, Hall C, Lim L (1990) The human heat-shock protein family. Expression of a novel heat-inducible HSP70 (HSP70B') and isolation of its cDNA and genomic DNA. Biochem J 267:125–132

Macario AJL, Dugan CB, Conway de Macario E (1991) A DnaK homologue in the archaebacterium *Methanosarcina mazei*. Gene 108:133–137

MacFarlane J, Blaxter ML, Bishop RP, Miles MA, Kelly JM (1990) Identification and characterization of a *Leishmania donovani* antigen belonging to the 70-kDa heat-shock protein family. Eur J Biochem 190:377–384

Maerz M, Rensing S, Igloi GL, Maier U-G (1992) Evolutionary analysis of the plastid-encoded gene for the α subunit of the DNA-dependent RNA polymerase of *Pyrenomonas salina* (Cryptophyceae). Curr Genet 22:479–482

Maier U-G (1992) The four genomes of the alga *Pyrenomonas salina* (Cryptophyta). Biosystems 28:69–73

Marshall JS, Keegstra K (1992) Isolation and characterization of a cDNA encoding the major HSP70 of the pea chloroplastic stroma. Plant Physiol 100:1048–1054

Milarski KL, Morimoto RI (1989) Mutational analysis of the human HSP70 protein: Distinct domains for nucleolar localization and adenosine triphosphate binding. J Cell Biol 109:1947–1962

Morimoto RI (1993) Chaperoning the nascent polypeptide chain. Curr Biol 3:101–102

Müller FW, Igloi GL, Beck CF (1992) Structure of a gene encoding heat shock protein HSP70 from the unicellular alga *Chlamydomonas reinhardtii*. Gene 111:165–173

Munro S, Pelham HRB (1986) An HSP70-like protein in the ER: Identity with the 78kd glucose-regulated protein and immunoglobulin heavy chain binding protein. Cell 46:291–300

Nicholson RC, Williams DB, Moran LA (1990) An essential member of the HSP70 gene family of *Saccharomyces cerevisiae* is homologous to immunoglobulin heavy chain binding protein. Proc Natl Acad Sci USA 87:1159–1163

Perkins LA, Doctor JS, Zhang K, Stinson L, Perrimon N, Craig EA (1990) Molecular and developmental characterization of the heat shock cognate 4 gene of *Drosophila melanogaster*. Mol Cell Biol 10:3232–3238

Roberts JK, Key JL (1991) Isolation and characterization of a Soybean *hsp*70 gene. Plant Mol Biol 16:671–683

Rochester DE, Winer JA, Shah DM (1986) The structure and expression of maize genes encoding the major heat shock protein, HSP70. EMBO J 5:451–458

Rzhetsky AR, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Saitou N, Imanishi T (1989) Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 6: 514–525

Scaramuzzi CD, Stokes HW, Hiller RG (1992) Heat shock HSP70 protein is chloroplast-encoded in the chromophytic alga *Pavlova lutherii*. Plant Mol Biol 18:467–476

Schwartz RM, Dayhoff MO (1978) Matrices for detecting distant relationships. In: Dayhoff MO (ed) Atlas of protein sequence and structure 1978. The National Biomedical Research Foundation, USA, pp 353–358

Slater MR, Craig EA (1989a) The SSA1 and SSA2 genes of the yeast *Saccharomyces cerevisiae*. Nucleic Acids Res 17:805–806

Slater MR, Craig EA (1989b) The SSB1 heat shock cognate gene of the yeast *Saccharomyces cerevisiae*. Nucleic Acids Res 17:4891

Ting J, Lee AS (1988) Human gene encoding the 78000-dalton glucose-regulated protein and its pseudogene: Structure, conservation, and regulation. DNA 7:275–286

Wang S, Liu XQ (1991) The plastid genome of *Cryptomonas* Φ encodes an HSP70-like protein, a histone-like protein, and an acyl carrier protein. Proc Natl Acad Sci USA 88:10783–10787

Watts FZ, Walters AJ, Moore AL (1992) Characterization of PHSP1, a cDNA encoding a mitochondrial *hsp*70 gene from *Pisum sativum*. Plant Mol Biol 18:23–32

Wetzstein M, Dedio J, Schumann W (1990) Complete nucleotide sequence of *Bacillus subtilis dnaK* gene. Nucleic Acids Res 18:2172

Winter J, Wright R, Duck N, Gasser C, Fraley R, Shah D (1988) The inhibition of Petunia *hsp*70 mRNA processing during $CdCl_2$ stress. Mol Gen Genet 211:315–319

Wu CFJ (1986) Jackknife, bootstrap and other resampling plans in regression analysis. Ann Statist 14:1261–1295