JOURNAL OF MOLECULAR EVOLUTION

# Heat Shock Protein 70 Family: Multiple Sequence Comparisons, Function, and Evolution

**Samuel Karlin, Luciano Brocchieri**

Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA

Received: 29 January 1998 / Accepted: 14 May 1998

**Abstract.** The heat shock protein 70 kDa sequences (HSP70) are of great importance as molecular chaperones in protein folding and transport. They are abundant under conditions of cellular stress. They are highly conserved in all domains of life: Archaea, eubacteria, eukaryotes, and organelles (mitochondria, chloroplasts). A multiple alignment of a large collection of these sequences was obtained employing our symmetric-iterative ITERALIGN program (Brocchieri and Karlin 1998). Assessments of conservation are interpreted in evolutionary terms and with respect to functional implications. Many archaeal sequences (methanogens and halophiles) tend to align best with the Gram-positive sequences. These two groups also miss a signature segment [about 25 amino acids (aa) long] present in all other HSP70 species (Gupta and Golding 1993). We observed a second signature sequence of about 4 aa absent from all eukaryotic homologues, significantly aligned in all prokaryotic sequences. Consensus sequences were developed for eight groups [Archaea, Gram-positive, proteobacterial Gram-negative, singular bacteria, mitochondria, plastids, eukaryotic endoplasmic reticulum (ER) isoforms, eukaryotic cytoplasmic isoforms]. All group consensus comparisons tend to summarize better the alignments than do the individual sequence comparisons. The global individual consensus ‘‘matches’’ 87% with the consensus of consensuses sequence. A functional analysis of the global consensus identifies a (new) highly significant mixed charge cluster proximal to the carboxyl terminus of the sequence highlighting the hypercharge run EEDKKRRER (one-letter aa code used). The individual Archaea and Gram-positive sequences contain a corresponding significant mixed charge cluster in the location of the charge cluster of the consensus sequence. In contrast, the four Gram-negative proteobacterial sequences of the alignment do not have a charge cluster (even at the 5% significance level). All eukaryotic HSP70 sequences have the analogous charge cluster. Strikingly, several of the eukaryotic isoforms show multiple mixed charged clusters. These clusters were interpreted with supporting data related to HSP70 activity in facilitating chaperone, transport, and secretion function. We observed that the consensus contains only a single tryptophan residue and a single conserved cysteine. This is interpreted with respect to the target rule for disaggregating misfolded proteins. The mitochondrial HSP70 connections to bacterial HSP70 are analyzed, suggesting a polyphyletic split of *Trypanosoma* and *Leishmania* protist mitochondrial (Mt) homologues separated from Mt-animal/fungal/plant homologues. Moreover, the HSP70 sequences from the amitochondrial *Entamoeba histolytica* and *Trichomonas vaginalis* species were analyzed. The *E. histolytica* HSP70 is most similar to the higher eukaryotic cytoplasmic sequences, with significantly weaker alignments to ER sequences and much diminished matching to all eubacterial, mitochondrial, and chloroplast sequences. This appears to be at variance with the hypothesis that *E. histolytica* rather recently lost its mitochondrial organelle. *T. vaginalis* contains two HSP70 sequences, one Mt-like and the second similar to eukaryotic cytoplasmic sequences suggesting two diverse origins.

**Key words:** Chaperonine — 70-kD heat shock protein — Archaea — Eubacteria — Eukaryotes — Evolution — SSPA — Multiple alignment

## Introduction

The heat shock protein 70 family (HSP70) is broadly and highly conserved across prokaryotes and eukaryotes. These proteins function for facilitating the assembly of multimeric protein complexes and as molecular chaperons for facilitating intracellular folding of proteins, for secretion and transport. They are activated rapidly for protecting the cell from environmental stress (Gething and Sambrook, 1992; Gupta and Golding, 1993; Boorestein et al. 1994; Gupta et al. 1997). Under normal cell conditions they constitute important and abundant proteins. In most eukaryotes from yeast to human there are at least three forms, with cellular location in the endoplasmic reticulum (ER) lumen, in the cytoplasm, and in mitochondrial and chloroplast organelles. We use the notation ER-euk and CYT-euk for the HSP70 sequences of the eukaryotic ER and cytoplasmic localizations, respectively. Several prokaryotic genomes contain multiple HSP70 sequences (see below).

This paper analyzes a large collection of HSP70 sequences among prokaryotes and eukaryotes. Our analysis centers on two protocols: SSPA (significant segment pair alignment) comparisons and development of a multiple alignment for all sequences. The SSPA method produces an assessment of the similarity for each pair of sequences (for formal details see Karlin et al. 1995; Brocchieri and Karlin 1998). The SSPA determinations distinguish consistent groups. Criteria for group ascertainments are twofold: (i) within-group SSPA scores generally exceed SSPA scores with sequences not in the group and (ii) SSPA scores with other groups or singular sequences are reasonably congruent for all members of the group. Applying our new method of multiple sequence alignment (Brocchieri and Karlin, 1998), ITERALIGN, we will establish a global alignment and separate alignments for subfamilies of the HSP70s: from archaeal sources, Gram-positive homologues, Gram-negative proteobacterial sequences, singular bacterial types, mitochondrial group, chloroplast functioning, and the two groups of eukaryotic isoforms ER-euk and CYT-euk separately. The consensus of these consensuses compares excellently (more than 86% identity) with the global consensus of the HSP70 individual sequences (see below).

## SSPA Assessments Among HSP70 Sequences

Pairwise SSPA scores (see Karlin et al. 1995; Brocchieri and Karlin 1998) were determined for more than 50 sequences. Following our protocol, a single representative sequence was retained from sets of sequences registering SSPA mutual scores exceeding 80%. By this criterion the HSP70 family was reduced to 40 sequences. Among prokaryotes the listing in Fig. 1 encompasses five archaeal sequences, six Gram-positive [Gram (+)] and one *Mycoplasma*; four Gram-negative proteobacterial sequences [Gram (−)]; a singular group consisting of two diverse *Synechococcus*, strains 1 and 2; and three other nonclassical individual eubacterial sequences. Among eukaryotes we have six mitochondria (Mt), one Mt-like sequence from the amitochondriate *Trichomonas vaginalis*, and two plastids (Pl), three ER-euk, and seven CYT-euk sequences. The protein sizes are reasonably constant ranging between 595 and 690 residues (two anomalous cases, lengths 706 and 749). No eocyte (*Sulfolobus*-like) HSP70 archaeal sequence has been detected to date.

*Archaeal SSPA Comparisons.* The SSPA scores among the archaeal sequences omitting the two halophiles are 56–65 (see Fig. 1). The two methanogen sequences score 61 and the two halobacteria mutually score 76 but, compared with the other Archaea, 49–58. The notation Archaea/Gram(+) and values refers to all the pairwise SSPA scores of archaeal sequences versus Gram(+) sequences and similarly for other group pairings. METMA (see legend to Fig. 1) aligns best with Gram(+) sequences, producing SSPA scores in the range 61–66 (exception, 70 with *Clostridium*) >> (non-*Methano* Archaea)/Gram(+) 51–57. Comparisons of Archaea to the mycoplasma MYCGE yield SSPA values lowered to 46–52. Pairwise alignments to Gram(−) sequences give METMA/Gram(−) 55–64 >> (non-*Methano* Archaea)/Gram(−) 49–53 [differences for each individual Gram(−) sequence are 5–12 points]. The SSPA scores of Archaea with mitochondrial and chloroplast sequences are again significantly higher for the methanobacteria than for other Archaea.

*SSPA Values Among Eubacterial Sequences.* The Gram(+) separate a high G+C group $G^+ = $ (*M. leprae*, *S. coelicolor*) having mutual SSPA score 73. The low G+C Gram(+) constitute three groups: $G_1 = \{$*B. subtilis*, *L. lactis*$\}$, $G_2 = \{$*Clostridium acetobutylicum*$\}$, and $G_3 = \{$*Erysipelothrix rhusiopathie*$\}$ with respect to SSPA values.

The Gram(−) divide into ${G^-}_1 = \{$ECOLI, PSECE$\}$ usually associated with γ-proteobacteria and ${G^-}_2 = \{$RHIME, CAUCR$\}$ associated with α-proteobacteria. The SSPA value within ${G^-}_1$ is 73 and that within ${G^-}_2$ is 76. Between-group alignment scores yield ${G^-}_1/{G^-}_2$ 67–69, consistent with their distinct proteobacterial designations.

Among all the classical eubacteria we have Gram(−)/Gram(+) 52–64 (mostly 57–64), with the highest scores achieved for *B. subtilis* and *Clostridium* aligned to ${G^-}_2$.

*SSPA Values for Mitochondrial (Mt) and Plastid (Pl) HSP70 Sequences Compared with Eubacterial Sequences.* The six mitochondrial HSP70 sequences divide into the two groups $M_1 = $ (animal, insect, yeast, and

| | | | | | PROKARYOTES | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | EUKARYOTES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Archaea | | | | | Eubacteria | | | | | | | | | | | | | | | | Organelles | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | Gram(+) | | | | | | | Gram(-) | | | | Singular | | | | | Mt. | | | | | | | Pl. | | ER | | | Cyt | | | | | | |
| | | | | | | | | | | $G^+$ | | $G_1$ | | $G_2$ | $G_3$ | | $G_1^-$ | | $G_2^-$ | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | METMA | METTH | THEAC | HALMA | HALCU | MYCLE | STRCO | BACSU | LACLA | CLOAC | ERYRH | MYCGE | ECOLI | PSECE | RHIME | CAUCR | BORBU | CHLTR | SYN1 | SYN2 | THEAQ | YEAST-Mt | DROME-Mt | MOUSE-Mt | PEA-Mt | TRYCR-Mt | LEIMA-Mt | TRIVA-Mt | PORPU-Pl | PEA-Pl | YEAST-ER | HUMAN-ER | GIALA-ER | GIALA | TRIVA | PLAFA | ENTHI | MAIZE | YEAST | HUMAN |
| | | | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| PROKARYOTES | Archaea | | | 1. METMA | - | 61 | 56 | 58 | 55 | 61 | 62 | 66 | 64 | 70 | 63 | 52 | 59 | 55 | 64 | 60 | 61 | 58 | 56 | 64 | 54 | 53 | 56 | 57 | 57 | 51 | 50 | 53 | 60 | 57 | 51 | 49 | 44 | 45 | 44 | 48 | 47 | 46 | 49 | 48 |
| | | | | 2. METTH | | - | 65 | 53 | 55 | 55 | 57 | 60 | 57 | 61 | 54 | 51 | 56 | 52 | 57 | 58 | 58 | 55 | 51 | 59 | 55 | 53 | 54 | 56 | 57 | 51 | 47 | 52 | 55 | 57 | 46 | 46 | 42 | 42 | 40 | 42 | 42 | 41 | 44 | 45 |
| | | | | 3. THEAC | | | - | 50 | 49 | 51 | 53 | 53 | 52 | 57 | 52 | 46 | 52 | 50 | 52 | 53 | 53 | 54 | 50 | 57 | 50 | 50 | 50 | 50 | 52 | 49 | 46 | 47 | 56 | 54 | 45 | 48 | 41 | 41 | 41 | 41 | 42 | 40 | 47 | 44 |
| | | | | 4. HALMA | | | | - | 76 | 54 | 56 | 55 | 54 | 55 | 53 | 48 | 51 | 50 | 55 | 52 | 51 | 49 | 49 | 53 | 53 | 49 | 49 | 50 | 50 | 48 | 47 | 49 | 50 | 48 | 44 | 46 | 41 | 41 | 41 | 41 | 41 | 40 | 43 | 43 |
| | | | | 5. HALCU | | | | | - | 55 | 54 | 54 | 53 | 53 | 52 | 47 | 50 | 49 | 53 | 51 | 51 | 49 | 48 | 52 | 50 | 49 | 47 | 49 | 50 | 44 | 46 | 48 | 51 | 48 | 44 | 42 | 39 | 40 | 40 | 41 | 41 | 39 | 41 | 41 |
| | Eubacteria | Gram(+) | $G^+$ | 6. MYCLE | | | | | | - | 73 | 60 | 60 | 60 | 57 | 51 | 56 | 55 | 59 | 57 | 59 | 56 | 52 | 61 | 56 | 50 | 53 | 53 | 55 | 48 | 48 | 53 | 54 | 56 | 46 | 47 | 41 | 42 | 44 | 45 | 45 | 44 | 46 | 46 |
| | | | | 7. STRCO | | | | | | | - | 62 | 61 | 62 | 57 | 53 | 56 | 58 | 59 | 57 | 59 | 57 | 53 | 62 | 55 | 52 | 53 | 54 | 56 | 49 | 48 | 53 | 58 | 61 | 47 | 47 | 41 | 42 | 43 | 46 | 47 | 42 | 48 | 45 |
| | | | $G_1$ | 8. BACSU | | | | | | | | - | 77 | 68 | 70 | 56 | 59 | 54 | 64 | 62 | 61 | 59 | 53 | 63 | 54 | 57 | 54 | 56 | 56 | 54 | 49 | 53 | 59 | 57 | 49 | 48 | 42 | 43 | 44 | 45 | 45 | 47 | 48 | 48 |
| | | | | 9. LACLA | | | | | | | | | - | 66 | 68 | 56 | 57 | 57 | 62 | 59 | 60 | 58 | 52 | 61 | 53 | 54 | 53 | 53 | 58 | 49 | 47 | 51 | 56 | 57 | 49 | 50 | 42 | 44 | 43 | 45 | 42 | 43 | 45 | 44 |
| | | | $G_2$ | 10. CLOAC | | | | | | | | | | - | 64 | 54 | 62 | 57 | 63 | 63 | 65 | 59 | 54 | 66 | 55 | 56 | 55 | 56 | 57 | 51 | 49 | 52 | 57 | 61 | 50 | 49 | 45 | 44 | 45 | 48 | 47 | 46 | 51 | 49 |
| | | | $G_3$ | 11. ERYRH | | | | | | | | | | | - | 60 | 57 | 52 | 57 | 61 | 59 | 55 | 51 | 59 | 54 | 53 | 52 | 54 | 56 | 49 | 47 | 51 | 57 | 57 | 51 | 49 | 43 | 46 | 47 | 48 | 48 | 48 | 51 | 51 |
| | | | | 12. MYCGE | | | | | | | | | | | | - | 51 | 51 | 51 | 51 | 52 | 52 | 49 | 55 | 51 | 46 | 47 | 49 | 50 | 47 | 45 | 42 | 51 | 49 | 47 | 44 | 40 | 42 | 42 | 44 | 43 | 43 | 45 | 46 |
| | | Gram(-) | $G_1^-$ | 13. ECOLI | | | | | | | | | | | | | - | 73 | 68 | 67 | 62 | 58 | 50 | 61 | 55 | 58 | 61 | 60 | 61 | 54 | 53 | 56 | 57 | 52 | 47 | 48 | 44 | 45 | 44 | 45 | 47 | 44 | 48 | 46 |
| | | | | 14. PSECE | | | | | | | | | | | | | | - | 69 | 68 | 62 | 56 | 48 | 59 | 50 | 57 | 58 | 59 | 59 | 53 | 52 | 55 | 53 | 52 | 46 | 47 | 42 | 41 | 42 | 44 | 43 | 42 | 48 | 45 |
| | | | $G_2^-$ | 15. RHIME | | | | | | | | | | | | | | | - | 76 | 61 | 58 | 51 | 60 | 57 | 64 | 64 | 65 | 66 | 58 | 58 | 61 | 57 | 53 | 49 | 50 | 45 | 46 | 47 | 47 | 47 | 45 | 50 | 49 |
| | | | | 16. CAUCR | | | | | | | | | | | | | | | | - | 58 | 59 | 50 | 59 | 54 | 59 | 62 | 63 | 65 | 56 | 54 | 60 | 56 | 53 | 47 | 48 | 44 | 45 | 47 | 44 | 46 | 44 | 48 | 48 |
| | | Singular | | 17. BORBU | | | | | | | | | | | | | | | | | - | 61 | 55 | 64 | 56 | 58 | 58 | 61 | 62 | 54 | 52 | 55 | 61 | 58 | 48 | 51 | 44 | 45 | 45 | 46 | 49 | 46 | 49 | 49 |
| | | | | 18. CHLTR | | | | | | | | | | | | | | | | | | - | 50 | 59 | 58 | 55 | 53 | 56 | 58 | 50 | 48 | 53 | 57 | 54 | 42 | 45 | 41 | 41 | 43 | 43 | 42 | 42 | 44 | 44 |
| | | | | 19. SYN1 | | | | | | | | | | | | | | | | | | | - | 62 | 59 | 47 | 45 | 45 | 47 | 44 | 44 | 49 | 63 | 53 | 42 | 43 | 39 | 39 | 39 | 37 | 40 | 41 | 43 | 42 |
| | | | | 20. SYN2 | | | | | | | | | | | | | | | | | | | | - | 60 | 53 | 59 | 55 | 58 | 55 | 50 | 52 | 73 | 73 | 47 | 48 | 44 | 46 | 45 | 48 | 46 | 46 | 49 | 49 |
| | | | | 21. THEAQ | | | | | | | | | | | | | | | | | | | | | - | 53 | 50 | 53 | 54 | 49 | 49 | 48 | 57 | 59 | 47 | 46 | 44 | 45 | 44 | 46 | 45 | 45 | 48 | 47 |
| | Organelles | Mt. | | 22. YEAST-Mt | | | | | | | | | | | | | | | | | | | | | | - | 65 | 66 | 59 | 59 | 54 | 60 | 53 | 50 | 48 | 49 | 42 | 43 | 45 | 46 | 47 | 45 | 48 | 48 |
| | | | | 23. DROME-Mt | | | | | | | | | | | | | | | | | | | | | | | - | 75 | 60 | 56 | 56 | 60 | 53 | 48 | 44 | 46 | 42 | 42 | 42 | 43 | 44 | 44 | 48 | 47 |
| | | | | 24. MOUSE-Mt | | | | | | | | | | | | | | | | | | | | | | | | - | 61 | 57 | 58 | 61 | 54 | 47 | 46 | 48 | 44 | 42 | 43 | 44 | 46 | 45 | 49 | 48 |
| | | | | 25. PEA-Mt | | | | | | | | | | | | | | | | | | | | | | | | | - | 58 | 57 | 63 | 56 | 51 | 44 | 46 | 42 | 41 | 43 | 45 | 43 | 44 | 48 | 46 |
| | | | | 26. TRYCR-Mt | | | | | | | | | | | | | | | | | | | | | | | | | | - | 80 | 57 | 49 | 45 | 44 | 45 | 39 | 39 | 42 | 43 | 43 | 41 | 45 | 43 |
| | | | | 27. LEIMA-Mt | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 55 | 47 | 45 | 44 | 44 | 40 | 40 | 42 | 42 | 43 | 42 | 44 | 43 |
| | | | | 28. TRIVA-Mt | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 50 | 50 | 45 | 47 | 41 | 42 | 41 | 46 | 43 | 45 | 47 | 47 |
| | | Pl. | | 29. PORPU-Pl | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 70 | 47 | 49 | 44 | 41 | 42 | 47 | 44 | 45 | 50 | 47 |
| | | | | 30. PEA-Pl | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 42 | 45 | 42 | 39 | 40 | 42 | 41 | 43 | 48 | 44 |
| EUKARYOTES | | ER | | 31. YEAST-ER | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 67 | 57 | 51 | 56 | 58 | 60 | 58 | 64 | 63 |
| | | | | 32. HUMAN-ER | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 56 | 58 | 58 | 60 | 61 | 62 | 65 | 64 |
| | | | | 33. GIALA-ER | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 53 | 52 | 52 | 55 | 54 | 57 | 57 |
| | | Cyt | | 34. GIALA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 57 | 61 | 61 | 63 | 65 | 66 |
| | | | | 35. TRIVA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 65 | 65 | 65 | 67 | 66 |
| | | | | 36. PLAFA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 69 | 70 | 74 | 75 |
| | | | | 37. ENTHI | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 74 | 77 | 75 |
| | | | | 38. MAIZE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 73 | 77 |
| | | | | 39. YEAST | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - | 77 |
| | | | | 40. HUMAN | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | - |

**Fig. 1.** Pairwise SSPA similarities of 40 HSP70 sequences. This set includes five archaeal sequences (the sequence code used throughout the text follows) including two methanobacteria, *Methanosarcina mazei* (1, METMA) and *Methanobacterium thermoautotrophicum* (2, METTH); one thermophile *Thermoplasma acidophilum* (3, THEAC); and two halophiles, *Halobacterium marismortui* (4, HALMA) and *Halobacterium cutirubrum* (5, HALCU); two high G+C Gram-positive, *Mycobacterium leprae* (6, MYCLE) and *Streptomyces coelicolor* (7, STRCO); four low C+G Gram-positive, *Bacillus subtilis* (8, BACSU), *Lactococcus lactis* (9, LACLA), *Clostridium acetobutylicum* (10, CLOAC), *Erysipelothrix rhusiopathiae* (11, ERYRH), and separately, *Mycoplasma genitalium* (12, MYCGE); four Gram-negative proteobacterial sequences, *E. coli* (13, ECOLI), *Pseudomonas cepacia* (14, PSECE), *Rhizobium meliloti* (15, RHIME), *Caulobacter crescentus* (16, CAUCR); five individual bacterial sequences including *Borrelia burgdorferi* (17, BORBU), *Chlamydia trachomatis* (18, CHLTR), two *Synechococcus*, strains 1 (19, SYN1) and 2 (20, SYN2), and *Thermus aquaticus* (21, THEAQ). Among the eukaryotes we have six mitochondrial (Mt) sequences, *Saccharomyces cerevisiae* (22, YEAST-Mt), *Drosophila melanogaster* (23, DROME-Mt), 24, MOUSE-Mt, 25, PEA-Mt, and the two Protists, *Trypanosoma cruzi* (26, TRYCR-Mt) and *Leishmania major* (27, LEIMA-Mt) and one mitochondrial-like sequence from *Trichomonas vaginalis* (28, TRIVA-Mt); two chloroplast (Pl) sequences, *Porphyra purpurea* (29, PORPU-Pl) and 30, PEA-Pl; three eukaryotic sequences from the endoplasmic reticulum (ER-euk), *S. cerevisae* (31, YEAST-ER), 32, HUMAN-ER, and *Giardia lamblia* (33, GIALA-ER); and seven sequences of cytoplasmic localization (CYT-euk) from *Giardia lamblia* (34, GIALA), *Plasmodium falciparum* (35, PLAFA), *Entamoeba histolytica* (36, ENTHI), *Trichomonas vaginalis* (37, TRIVA), 38, MAIZE, 39, YEAST, and 40, HUMAN.

plant Mt) and $M_2$ = (protist Mt: TRYBR, LEIMA). The sequence from the amitochondriate *Trichomonas vaginalis* (TRIVA-Mt) is more similar to $M_1$ sequences (60–63) than to $M_2$ sequences (55, 57). Comparisons to the Gram(−) and Gram(+) sequences entail SSPA values Mt/Gram(−) 50–66. The highest Mt/bacterial scores are achieved with the Gram(−) α-proteobacterial types at the level 62–66 (one 59) for $M_1$ and 52–58 for $M_2$. An Mt-like sequence from the amitochondriate protist *Varimorpha necatrix* (Microsporidia) also shows the highest similarity to $M_1$ and α-proteobacterial sequences but with reduced SSPA scores, 51–52 (data not shown). The two Pl score high, with a SSPA value of 70, but lower than the alignment scores to SYN2 (score 73), consistent with the Pl endosymbiont hypothesis. Both the Mt and the Pl sequences compared with all eukaryotic sequences yield SSPA values in the relatively low range, 39–50.

*SSPA Scores Between ER-euk and CYT-euk HSP70 Sequences.* Three HSP70 sequences (yeast, *Giardia,* and human) have a subcellular localization to the ER. The SSPA score is relatively high at 67 for yeast compared to

```
                              1                    2                       3                        4                      5                 6                 7                                      8
   CONS.                    G-K-V  0 IGIDLGTTNSCVAVMEGGKPKVIANAEG  1 RTTPSVVAFT  1 DGERLVGDPAKRQAVTNPENTIFSVKRLIG  0 RRFDD  2 VQKDMKHVPYKV-V---KA-D  5 VEV  2 EG--KEYTPEEISAMVLQKMKETAESYLGEKVTNAVITVPAYFNDSQRQATKDAGKIAGLNVLRIINEPTAAALAYGL-DK  3

 1 METMA          2 A K I     LGIDLGTTNSCVAVMEGGEAVVIPNAEG  s RTTPSVVGFS  k KGEKLVGQVAKRQAISNPDNTVYSIKRHMG                                       5 VTL      NG   KDYTPQEISAMILQKLKADAEAYLGETIKQAVITVPAYFNDSQRQATKDAGAIAGLEVLRIINEPTAASLAYGL DK  3
 2 METTH          8 E K I     LGIDLGTSNSAAAVLIGGKPTIIPSAEG  5 KSFPSCVAFT  e DGQMLVGEPARRQAVTNPENTITAIKRSMG                                       4 VKV      HG   KEYTPQEISAFILQKIKKDAEAFLGEEIKKAVITVPAYFDDNQRTATKDAGTIAGLDVVRLVNEPTAASLAYGL DK  3
 3 THEAC   A      2 S K I     IGIDLGTSNSAAAVVISGKPTVIPSSEG  5 KAFPSYVAFT  k DGQMLVGEPARRQALLNPEGTIFAAKRKMG                                       4 FKV    f D    KEFTPQQISAFILQKIKKDAEAFLGEPVNEAVITVPAYFNDNQRQATKDAGTIAGFDVKRIINEPTAAALAYGV DK  4
 4 HALMA          4 N K I     LGIDLGTTNSAFAVMEGGDPEIIVNGEG  e RTTPSVVAFD    DGERLVGKPAKNQAVKNPDETIQSIKRHMG                                       5 VEL      DG   EEYTPEQVSAMILQKIKHDAEEYLGDEIEKAVITVPAYFNDRQRQATKDAGKIAGFEVERIVNEPTAAAMAYGL DD es
 5 HALCU          4 E K I     LGVDLGTTNSAFAVMEGSDPEIITNEEG  d LTTPSIVAHD    DGELLVGKPAKNQAVQNPDQTIASIKRHMG                                       5 VAL      GG   DEYTPEEISARILQKIKRDAEEYLGQDVEKAVITVPAYFNDRQRQATKDAGEIAGFDVERIVNEPTAASMAYGL DE dr
 6 MYCLE          2 A R A     VGIDLGTTNSVVSVLEGGDPVVVANSEG  s RTTPSTVAFA  r NGEVLVGQPAKNQAVTNVDRTIRSVKRHMG                                       4 IEI      DG   KKYTAQEISARVLMKLKRDAEAYLGEDITDAVITTPAYFNDAQRQATKEAGQIAGLNVLRIVNEPTAAALAYGL DK  3
 7 STRCO          2 A R A     VGIDLGTTNSVVSVLEGGEPTVITNAEG  a RTTPSVVAFA  k NGEVLVGEVAKRQAVTNVDRTIRSVKRHMG                                       4 VNL      DG   KDFNPQQISAFVLQKLKRDAESYLGEKVTDAVITVPAYFNDAERQATKEAGEIAGLNVLRIVNEPTAAALAYGL DK  d
 8 BACSU          1 S K V     IGIDLGTTNSCVAVLEGGEPKVIANAEG  n RTTPSVVAFK    NGERQVGEVAKRQSITNP NTIMSIKRHMG                                       4 VEI      EG   KDYTPQEVSAIILQHLKSYAESYLGETVSKAVITVPAYFNDAERQATKDAGKIAGLEVERIINEPTAAALAYGL DK  3
 9 LACLA   +      2 S K I     IGIDLGTTNSAVAVLEGTESKIIPNPEG  n RTTPSVVAFK    NGEIIVGDAAKRQAVTNPE TIISIKSKMG                                       4 VSA      NG   KEYTPQEISAMILQNLKATAESYLGEKVEKAVITVPAYFNDAQRQATKDAGKIAGLEVERIVNEPTAAALAYGL DK  3
10 CLOAC          2 S K V     IGIDLGTTNSCVAVMEGGDPAVIANSEG  a RTTPSVVSFQ  k NGERLVGQVAKRQSITNPDKTIISIKRKMG                                       4 VAI      DD   KNYTPQEISAMILQKLKADAEAYLGETVTQAVITVPAYFNDSQRQATKDAGKIAGLEVLRIINEPTAASLAYGL DK  3
11 ERYRH          2 S K V     IGIDLGTTNSAVSVMDGGEAKVITNPEG  n RTTPSVVSFK    NGERIVGDAAKRQVVTNP NSAVSVKRLIG                                       4 VTL      EG   KDYTPEEISAMILGYMKSYAEDYLGEKVTKAVITVPAYFNDAQRQATKDAGKIAGLEVERIINEPTAAALAFGI DK  4
12 MYCGE          6 G L I     IGIDLGTTNSCVSVMEGGRPVVLENPEG  k RTTPSIVSYK    NNEIIVGDAAKRQMVTNP NTIVSIKRLMG                                       4 VKV    3 DGttKELSPEQVSAQILSYLKDFAEKKIGKKISRAVITVPAYFNDAERNATKTAGKIAGLNVERIINEPTAAALAYGI DK  3
13 ECOLI          1 G K I     IGIDLGTTNSCVAIMDGTTPRVLENAEG  d RTTPSIIAYT  q DGETLVGQPAKRQAVTNPQNTLFAIKRLIG   RRFQD ee VQRDVSIMPFKI I   AA D   5 VEV      KG   QKMAPPQISAEVLKKMKKTAEDYLGEPVTEAVITVPAYFNDAQRQATKDAGRIAGLEVKRIINEPTAAALAYGL DK  3
14 PSECE   -      2 G K I     IGIDLGTTNSCVAIMEGNQVKVIENSEG  t RTTPSIIAYM  d DNEVLVGAPAKRQSVTNPKNTLFAVKRLIG   RRFEE ke VQKDIGLMPYSI I   KA D   5 VEG      HG  eKLMAPPQVSAEAVRKMKKTAEDYLGEPVTEAVITVPAYFNDSQRQATKDAGRIAGLEVKRIINEPTAAALAFGL DK  4
15 RHIME          2 A K V     IGIDLGTTNSCVSVMDGKDAKVIENAEG  a RTTPSMVAFT  e DGERLVGQPAKRQAVTNPENTLFAIKRLIG   RTFED pt TQKDKGMVPYKI V   KA D   5 VEA      HG   TSYSPSQISAMILQKMKETAESYLGEKVEKAVITVPAYFNDAQRQATKDAGKIAGLDVLRIINEPTAAALAYGL DK ke
16 CAUCR          2 S K I     IGIDLGTTNSCVAIMDGKTPKVIENAEG  a RTTPSVVAFL  e DGERLIGQPAKRQAVTNPTNTLFAIKRLIG   RTASD pv VEKDKGMVPYRS S   RA R   5 VKA      HG   KDYSPQEVSAFILQKMKEAAEAHLGEPVTKAVITVPAYFNDAQRQATKDAGKIAGLEVLRIINEPTAAALAYGL DK  n
17 BORBU          2 G K I     IGIDLGTTNSCVAIMEHGKPVVIQNSEG  g RTTPSIVAYT  n KGERLVGQVAKNQMVTNPENTIYSIKRFMG   RRFEE    VASEIKMVPYKI E   KG L   5 VNI    3 K    KQMSPPEISAATLTKMKETAEAYLGEKVTEAVITVPAYFNDAQRQATKDAGKIAGLEVKRIVNEPTAAALAYGI EK kh
18 CHLTR          7 N K I     IGIDLGTTNSCVSVMEGGQPKVIASSEG  t RTTPSIVAFK    GGETLVGIPAKRQAVTNPEKTLASTKRFIG   RKFSE    VESEIKTVPYKV ApnsKG D  av FDV      EQ   KLYTPEEIGAQILMKMKETAEAYLGETVTEAVITVPAYFNDSQRASTKDAGRIAGLDVKRIIPEPTAAALAYGI DK eg
19 SYN1    S      2 G R V     VGIDLGTTNSVIAVMEGGKPMVIANAEG  v RTTPSVVGVS  k TGERLVGELARRQLVLNPRNTFANIKRFIG   RRYDE    LTDESKRVPY T V   RR D   4 VRI    4 LS   REFAPEEVAAMILRKLAEEASRYLGEPVTGAVITVPAYFNDSQRQATRDAGRIAGLEVKRILNEPTAASLAYGL DR rd
20 SYN2           2 A K V     VGIDLGTTNSCVAVMEGGKPTVIANAEG  f RTTPSVVAFA  k NQDRLVGQIAKRQAVMNPENTFYSVKRFIG   RRPDE    VTNELTEVAYKVdT   SG N   a VKL    4 AG   KQFAPEEISAQVLRKLAEDASKYLGETVTQAVITVPAYFNDSQRQATKDAGKIAGLEVLRIINEPTAAALAYGL DK ks
21 THEAQ          2 A K A     VGIDLGTTNSVIAVLEGGKPVVLENAEG  e RVTPSVVAFR    DGETLVGRMAKRQAVLNPEGTIFEIKRFIG   RRFEE    VQEEAKRVPYKV V   PGpD   4 VEV      KG   KLYTPEEISAMILRKLVEDASKKLGEKITKAVITVPAYFNNAQREATANAGRIAGLEVLRIINEPTAAALAYGL DK kg
22 YEAST         30 G S V     IGIDLGTTNSAVAIMEGKVPKIIENAEG  s RTTPSVVAFT  k EGERLVGIPAKRQAVVNPENTLFATKRLIG   RRFED ae VQRDIKQVPYKI V   KH S   5 VEA      RG   QTYSPAQIGGFVLNKMKETAEAYLGKPVKNAVVTVPAYFNDSQRQATKDAGQIVGLNVLRVVNEPTAAALAYGL EK  s
23 DROME         48 G E V   4 IGIDLGTTNSCLAVMEGKQAKVIENAEG  a RTTPSHVAFT  k DGERLVGMPAKRQAVTNSANTFYATKRLIG   RRFDD pe VKKDITNLSYKV V   KA S   5 VSS    t DG   KVYSPSQIGAFILMKMKETAEAYLNTPVKNAVVTVPAYFNDSQRQATKDAGQIAGLNVLRVINEPTAAALAYGM DK te
24 MOUSE         53 G A V     VGIDLGTTNSCVAVMEGKQAKVLENAEG  a RTTPSVVAFT  a DGERLVGMPAKRQAVTNPNNTFYATKRIIG   RRYDD pe VQKDTKNVPFKI V   RA S   5 VEA      HG   KLYSPSQIGAFVLMKMKETAENYLGHTAKNAVITVPAYFNDSQRQATKDAGQISGLNVLRVINEPTAAALAYGL DK se
25 PEA     Mt    51 G NdV     IGIDLGTTNSCVSVMEGKNPKVIENSEG  a RTTPSVVAFN  q KSELLVGTPAKRQAVTNPTNTLFGTKRLIG   RRFDD aq TQKEMKMVPYKI V   RA P   5 VEA      NG   QQYSPSQIGAFVLTKIKETAEAYLGKTISKAVVTVPAYFNDAQRQATKDAGRIAGLDVQRIINEPTAAALSYGMnNK
26 TRYCR         27 G D V     IGIDLGTTYSCVAVMEGDKPRVLENTEG  f RATPSVVAFK    GQEKLVGLAAKRQAVTNPQSTFFAVKRLIG   RRFED sn IQHDIKNVPYKI G   RS S   5 VQD    a NG   KQYSPSQVGAFVLEKMKETAENFLGRKVSNAVVTCPAYFNGPQRQATKDAGTIAGLNVIRVVNGPTAAALAYGL DK tk
27 LEIMA         27 G D V     IGVDLGTTYSCVATMDGDKARVLENSEG  f RTTPSVVAFK    GSEKLVGLAAKRQAITNPQSTFYAVKRLIG   RRFED eh IQKDIKNVPYKI V   RA G   5 VQD    g NG   KQYSPSQIGAFVLEKMKETAENFLGHKVSNAVVTCPAYFNDAQRQATKDAGTIAGLNVIRVVNEPTAAALAYGM DK tk
28 TRIVA          5          IGIDLGTTNSCVSVVEGGTPKVIQNAEG  v RTTPSIVAFT  n TGERLVGEQAKRQAITNSKSTFFATKRLIG   CSFDD em TKKCREMVPYQI V   KA K   5 VKD    e KG   NEYSPSQIGAFVLMKMKETAEAYLGTSVKDAVITVPAYFNDSQRQATKDAGKIAGLNVMRTLNEPTAAALAYGT ER kq
29 PORPU   Pl     2 G K V     VGIDLGTTNSVIAVMEGGKPTVIPNAEG  f RTTASVVAYT  k SGDKLVGQIA RQAVINPENTFYSVKRFIG 5          ISQEIRQTSYN  V   KT S   3 IKI    4 LN   KDFAPEEISAQVLRKLVEDASTYLGETVTQAVITVPAYFNDSQRQATKDAGKIAGLDVLRIINEPTAASLSYGL DK qn
30 PEA           68 E K V     VGIDLGTTNSAVAAMEGGKPTIITNAEG  q RTTPSVVAYT  k NGDRLVGQIAKRQAVVNPENTFFSVKRFIG   RKMSE    VDEESKQVSYRV I   RD D   3 VKL    4 IG   KSFAAEEISAQVLRKLVDDASKFLNDKVTKAVVTVPAYFNDSQRTATKDAGRIAGLEVLRIINEPTAASLAYGF ER kn
31 YEAST         50 G T V     IGIDLGTTYSCVAVMKNGKTEILANEQG  n RITPSYVAFT    DDERLIGDAAKNQVAANPQNTIFDIKRLIG   LKYND rs VQKDIKHLPFNV V   NK D   4 VEV    4 EK   KVFTPEEISGMILGKMKQIAEDYLGTKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVLRIVNEPTAAAIAYGL DK  3
32 HUMAN   ER    28 G T V     VGIDLGTTYSCVGVFKNGRVEIIANDQG  n RITPSYVAFT  p EGERLIGDAAKNQLTSNPENTVFDAKRLIG   RTWND ps VQQDIKFLPFKV V   EK K   4 IQV    5 QT   KTFAPEEISAMVLTKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGL DK  3
33 GIALA          3 A L V  10 IGIDLGTTYSCVAVSRAGQVEIIPNELG  a RVTPSYVAFT  a DGERLVGDAAKNYAPISPENTIFDVKRLIG   RKFDD pe VQKDMKLLPYKV I   NK D   4 VQL    9 QN   KIMSPEEISAMVLTKMKTIAEDYLGEKITKAVVTVPAYFSDSQRSATKDAGRIAGLDVVRIINEPTSSSIAYGL DK  8
34 GIALA          7          VGIDLGTTYSCVGVYQNEKVEIIANEQG  a YTTPSYVAFT    DADGLIGDSAKNQCALNPENTIFDAKRLIG   RRFND pe VQADLKH  FR  S   RS S   8 IQV    4 ET   KTFTPEEISSMVLTKMKDIASDYLGNKVTEAIVTVPAYFSDAHGKATQNAGTIAGLNVLRIINEPTAAAIAYGL DK  5
35 TRIVA          4          IGIDLGTTYSCVGVFRHNRVDILANDMG  n RTTPSYVAFT  e D ERLVGDAAKNQAAMNPQNTVFDAKRMIG   RKFSD pv VQKDMKHWPFKV V   RS D   5 IEV    4 ET   KQFFPEEISAMVLGKMKETAENMLSEKITDAVVTVPAYFNDGQRQATKDAGVIAGLNVLRIINEPTAAAIAFGL NE  4
36 PLAFA          6 GsK P   8 IGIDLGTTYSCVGVWRNENVDIIANDQG  n RTTPSYVAFT    DTERLIGDAAKNQVARNPENTVFDAKRLIG   RKFTE ss VQSDMKHWPF T V   KS G   6 IEV    4 EK   KLFHPEEISSMVLQKMKENAEAFLGKSIKNAVITVPAYFNDSQRQATKDAGTIAGLNVMRIINEPTAAAIAYGL HK  4
37 ENTHI   CYT    6 G P A     VGIDLGTTYSCVGIWQNDRVEIIANDQG  n RTTPSYVAFT    DTERLIGDAAKNQIAMNVKNTVFDAKRLIG   RRFSD pa IQNDMKHWSFKV I   D  D   6 IEV    4 EV   KKFTPEEISSMVLTKMKETAESFVGKEVKNAVITCPAYFNDSQRQATKDAGTIAGMNVMRIINEPTAAAIAYGL DK  4
38 MAIZE          6 G P A     IGIDLGTTYSCVGLWQHDRVEIIANDQG  n RTTPSYVGFT    DTERLIGDAAKNQVAMNPTNTVFDAKRLIG   RRFSS pa VQSSMKLWPSRH L  gLG D   3 I V    5 EE   KQFAAEEISSMVLIKMKEIAEAYLGSTIKNAVVTVPAYFNDSQRQATKDAGVIAGLNVMRIINEPTAAAIAYGL DK  6
39 YEAST          1 S K A     VGIDLGTTYSCVAHFANDRVDIIANDQG  n RTTPSFVAFT    DTERLIGDAAKNQAAMNPSNTVFDAKRLIG   RNFND pe VQADMKHFPFKL I   DV D   4 IQV    4 ET   KNFTPEQISSMVLGKMKETAESYLGAKVNDAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAAIAYGL DK  3
40 HUMAN          4 G P A     VGIDLGTTYSCVGVFQHGKVEIIANDQG  n RTTPSYVAFT    DTERLIGDAAKNQVAMNPTNTVFDAKRLIG   RRFDD av VQSDMKHWPFMV V   N  D   5 VQV    4 ET   KSFYPEEVSSMVLTKMKEIAEAYLGKTVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAAIAYGL DK  4
                        *     *1*1111**1***** *    ** * *1     ****1 ***       * **1  1***** ** ***   1* *1    *** *    ** * *- ****               *        *-* * **** ** ***  1* ***  *  1**1*1111*******1**11 1*1**1 1****11******1* **


                            9                              10                                  11                              12            13                           14                                   15
   CONS.               D-KTILVFDLGGGTFDVSILEI  0 GD-GVFEVKATNGDTHLGGEDFDNRIVDYLVEEFKKEN  0 GIDLSKD---KMALQRLREAAEKAKIELSSTTQ-T  1 INLP-FITA-G-A  1 G-PKH  0 LEMTLTRAKFEELTEDLIER-TLGPVEQALKDAGLSKSDIDEVILVGGSTRMP  1 VQQLVKDFF  0

 1 METMA        164 D QKILVYDLGGGTFDVSILEL     GG GVFEVKSTSGDTHLGGDDFDQRVIDYLLAEFKKSE     GIDLSKD   KAVLQRLKDAAEKAKIELSGVAN T  n INLP FLTV G T   d GePKH    MDIDLTRAQFQKMTEDLLEK TLVSMRRALSDAKLTPNDLDKVILVGGATRMP  a VVELVENFT
 2 METTH        173 D MVIMVFDLGGGTLDVTIMEF     GG GVFEVRSTSGDTQLGGTDMDNAIMNYLAEEFKMET     GIDLMED   DQAVQRLREAAEKAKIEL STTL T te VNLP YITV   A  qd G PKH    LIKTITRAKLEELVDPIVQK CAGPMEQALRDAGMTREDVDKIILVGGPTRMP  i VQKFVEDFI
 3 THEAC   A    168 E K ILVFDLGGGTLDVTIMDF     GD AVFQVLSTSGDTRLGGTDMDEAIVNYIADDFQKKE     GIDLRKD   RSAYIRLRDAAEKAKIELSTTLS T  d IDLP YITVtN S      G PKH    IKMTLTRAKLEELYSPIVER VKGPIDKALEGAKLKKTEITKLLFVGGPTRIP  y VRKYVEDYL
 4 HALMA        164 D QTVLVYDLGGGTFDVSILDL     GG GVYEVVATNGDNDLGGDDWDHAIIDYLADEFEAEH     GIDLRDD   RQALQRLTEAAEEAKIELSSRKE T  r INLP FI A T T  dd G PLD    LEQKITRAKFESLTEDLIER TLGPTEQALADADYTKSDIDEVILVGGSTRMP  q VQDQVEEMT
 5 HALCU        164 D QTVLVYDLGGGTFDVSILDL     GG GVYEVAATNGDNDLGGDDWDHAIIDHLADNFENEH     GIDLRED   RQALQRLTEAAEEAKIELSSRKE T  t VNLP FVTAtD S      G PVH    LEQDITRATFESITEDLIER TVGPTEQALEDAGLSKSDIDDVILVGGSTRMP  q VQAQVEDLV
 6 MYCLE        163 E QTILVFDLGGGTFDVSLLEI     GE GVVEVRATSGDNHLGGDDWDDRIVNWLVDKFKGTS     GIDLTKD   KMAMQRLREAAEKAKIELSSSQS T  s VNLP YITV D S  dk N PLF    LDEQLIRAEFQRITQDLLDR TRQPFQSVVKDAGISVSEIDHVVLVGGSTRMP  a VTDLVKELT  g
 7 STRCO        161 E QVILVFDLGGGTFDVSLLEI     GD GVVEVKATNGDNNLGGDDWDQRVVDYLVKQFQSGH     GVDLAKD   KMALQRLREAAEKAKIELSSSTE T  t INLP YITA S A   e G PLH    LDEKLTRAQFQQLTSDLLER CKTPFHNVIKDAGIQLSEIDHVVLVGGSTRMP  a VAELVKELT  g
 8 BACSU        160 D QTILVYDLGGGTFDVSILEL     GD GVFEVRSTAGDNRLGGDDFDQVIIDHLVSEFKKEN     GIDLSKD   KMALQRLKDAAEKAKKDLSGVSS T  q ISLP FITA GeA      G PLH    LELTLTRAKFEELSSHLVER TMGPVRQALQDAGLSASEIDKVILVGGSTRIP  a VQEAIKKET
 9 LACLA   +    161 D EKILVFDLGGGTFDVSILEL     GD GVFDVLATAGNNKLGGDDFDQKIIDWMVAEFKKEN     GIDLGQD   KMALQRLKDAAEKAKKDLSGVTT T  q ISLP FITA G A   a G PLH    LEMALTRAKFDELTHDLVEA TRQPVRQALSDAGLSTSDIDEVLLVGGSTRIP  a VVELVRHET
10 CLOAC        163 N QKILVYDLGGGTFDVSVLEL     GD GVFEVKSTNGNTHLGGDDFDEKIMDYIAEEFKKDN     GIDLRND   KMALQRLKEAAEKAKIELSSSTQ T  n INLP FITA D A   t G PKH    IDMNLTRAKFNELTEGLVQD TIEPMKKALSDAGLSINDIDKIVLVGGSTRIP  a VQEAVKNYT
11 ERYRH        162 E K VLVFDLGGGTFDVSILEL     AD GTFEVLSTAGDNKLGGDDFDNIVVDYLVDIFKKEN     GIDLSSD   KMAMQRLKEAAEKAKKDLSSTVN A  s ISLP FISA G E   n G PLH    LETTLSRAKFEEMTKSLVER TMVPVRQALKDAGLTKNDIHQVLLVGGSTRIP  a VVEAVKNDL
12 MYCGE        170 E MKVLVYDLGGGTFDVSLLDI     AE GTFEVLATAGDNRLGGDDWDNKIIEYISAYIAKEH   q GLNLSKD   KMAMQRLKEAAERAKIELSAQLE T  i ISLP FLTV T Q   k G PVN    VELKLTRAKFEELTKPLLER TRNPISDVIKEAKIKPEEINEILLVGGSTRMP  a VQKLVESMV  p
13 ECOLI        186 N RTIAVYDLGGGTFDISIIEI   4 GE KTFEVLATNGDTHLGGEDFDSRLINYLVEEFKKDQ     GIDLRND   PLAMQRLKEAAEKAKIELSSAQQ T  d VNLP YITA D A   t G PKH    MNIKVTRAKLESLVEDLVNR SIEPLKVALQDAGLSVSDIDDVILVGGQTRMP  m VQKKVAEFF
14 PSECE   -    189 D RKIAVYDLGGGTFDVSIIEI   4 GE MQFEVLSTNGDTFLGGEDFDQRIIDYIIGEFKKEQ     GVDLSKD   VLALQRLKEAAEKAKIELSSSQQ T  e INLP YITA D A   s G PKH    LNLKITRAKLEALVEDLVER TIEPCRIAIKDAGVKVSDIDDVILVGGQTRMP  k VMEKVKEFF
15 RHIME        186 G KTIAVYDLGGGTFDISVLEI     GD GVFEVKSTNGDTFLGGEDFDMRLVEYLASEFKKEQ     GIDLKND   KLALQRLKEAAEKAKIELSSSQQ T  e INLP FITA D A   s G PKH    LTMKLSRAKFESLVEDLIQK TIAPCKAALKDAGVSAAEIDEVVLVGGMTRMP  k VQETVKQLF
16 CAUCR        185 DgKKIAVYDLGGGTFDVSILEI     GD GVFEVKSTNGDTFLGGEDFDLRIVDYLADEFKKEQ     GVDLRKD   KLALQRLRE AEKAKKELSSTAQ Y  e VNLP FISM N A   s G PLH    LNIKLSRAKLEALVDDLIAR TIGPCEQALKDAGLKKSDIDEVILVGGMSRMP  k VQQAVQDFF
17 BORBU        186 E EIVAVYDLGGGTFDISILEL     GD GVFEVKSTNGDTHLGGDNFDDEIIKHLISEFKKES     AIDLSND   KMALQRLKEAAEKAKIELSGAQE A  s INLP FITA D A   n G PKH    LQYTLTRAKFEQMVDHLVQK TKEPCLKAIKDAGLKASDINEVILVGGSTRIP  a IQKIVKDIF
18 CHLTR        188 D KKIAVFDLGGGTFDISILEI     GD GVFEVLSTNGDTHLGGDDFDGVIINWMLDEFKKQE     GIDLSKD   NMALQRLKDAAEKAKIELSGVSS T  e INQP FITI D A   n G PKH    LALTLTRAQFEHLASSLIER TKQPCAQALKDAKLSASDIDDVLLVGGMSRMP  a VQAVVKEIF
19 SYN1    S    186 N QTILVFDLGGGTFDVSVLKV     GN GVFEVKATSGDTQLGGNDFDRRIVDWLAEQFLEAE     GIDLRRD   RQALQRLIEAAEKAKIELSGVSV T  d INLP FITA T E   d E PKH    LETRLTRSEFEALCEDLLER MMRPLRRALKDARLQPQDIDEVVLVGGSTRMP  m VQQLVRSLI
20 SYN2         185 N ERILVFNLGGGTFDVSVLEV     GD GVFEVLATSGDTHLGGDDFDKKIVDFLAGEFQK N   e GIDLRKD   KQALQRLTEAAEKAKIELSSATQ T  e INLP FITA T Q   d G PKH    LDLTLTRAKFEELASDLIDR CRIPVEQAIKDAKLALSEIDEIVLVGGSTRIP  a VQAIVKQMT
21 THEAQ        183 N ETVLVFDLGGGTFDVTILEI     GE GVFEVKATSGDTHLGGSDMDHAIVNWLAEEFKKEH     GVDLKAD   RQALQRLIEAAEKAKIELSSTLE T  t ISLP FIAL DpA  sk T PLH    LEKKLTRAKFEELIQPLLKR LRGPVEQALKDAGLTPAQIDEVILVGGATRVP  a VQQVVRELL
22 YEAST        213 DsKVVAVFDLGGGTFDISILDI     DN GVFEVKSTNGDTHLGGEDFDIYLLREIVSRFKTET     GIDLEND   RMAIQRIREAAEKAKIELSSTVS T  e INLP FITA D A   s G PKH    INMKFSRAQFETLTAPLVKR TVDPVKKALKDAGLSTSDISEVLLVGGMSRMP  k VVETVKSLF
23 DROME        237 D KIIAVYDLGGGTFDISILEI     QK GVFEVKSTNGDTLLGGEDFDNHIVNFLVVEFKKDS     GIDIRKD   NIAMQRLKEAAEKAKCELSSSQQ T  d INLP YLTM D A   a G PQH    MNLKLTRSKLESLVGDLIKR TIQPCQKALSDAEVSKSEIGEVLLVGGMTRMP  k VQSTVQELF
24 MOUSE        237 D KVIAVYDLGGGTFDISILEI     QK GVFEVKSTNGDTFLGGEDFDQALLRHIVKEFKRET     GVDLTKD   NMALQRVREAAEKAKCELSSSVQ T  d INLP YLTM D A   s G PKH    LNMKLTRAQFEGIVTDLIKR TIAPCQKAMQDAEVSKSDIGEVILVGGMTRMP  k VQQTVQDLF
25 PEA     Mt   235 E GLIAVFDLGGGTFDVSILEI     SN GVFEVKATNGDTFLGGEDFDNALLDFLVSEFKRTE     SIDLAKD   KLALQRLREAAEKAKIELSSTSQ T  e INLP FISA D A   s G AKH    LNITLTRSKFEALVNNLIER TKAPCKSCLKDANISIKDVDEVLLVGGMTRVP  k VQQVVSEIF
26 TRYCR        211 D SMIAVYDLGGGTFDISVLEI   a G  GVFEVKATNGDTHLGGEDFDLCLSDYILTEFKKST     GIDLSNE   RMALQRIREAAEKAKCELSTTME T  e VNLP FITA N Q   d G AQH    VQMTVSRSKFESLAEKLVQR SLGPCKQCIKDAAVDLKEISEVVLVGGMTRMP  k VIEAVKQFF
27 LEIMA        211 D SLIAVYDLGGGTFDISVLEI   a G  GVFEVKATNGDTHLGGEDFDLALSDYILEEFRKTS     GIDLSKE   RMALQRVREAAEKAKCELSSAME T  e VNLP FITA N A   d G AQH    IQMRISRSKFEGITQRLIER SIAPCKQCMKDAGVELKEINDVVLVGGMTRIR  9
28 TRIVA        187 G QTVAVFDLGGGTFDFSILEI   s KD GVFEVKATNGDTFLGGEDFDAALMKYVIQDFQAKN     QIDLAKD   PLALQRIREAVEKAKCELSSMMT T  e INLP YITVtG A      G PKH    LQMPITRATFEKITQHLIAR TINPCKNCLKDAGLTPQQINEVILVGGMTRMP  k VIDSVKEFF
29 PORPU   Pl   184 N ETILVFDLGGGTFDVSILEV     GD GVFEVLSTSGDTHLGGDDFDQQIVEWLIKDFKQSE     GIDLGKD   RQALQRLTEASEKAKIELSNLTQ T  e INLP FITA T Q   d G PKH    LEKTVTRAKFEELCSRLIDKcSI PVNNALKDAKLEASSIDEVVLVGGSTRIP  a IQQMVKRLI
30 PEA          252 N ETILVFDLGGGTFDVSVLEV     GD GVFEVLSTSGDTHLGGDDFDKRVVDWLAGDFKRDE     GIDLLKD   KQALQRLTETAEKAKMELSSLSQ T  n ISLP FITA T A   d G PKH    IETTLTRAKFEELCSDLLDR LRTPVENSLRDAKLSIKDIDEVILVGGSTRIP  a VQELVKKLI
31 YEAST        237 E HQIIVYDLGGGTFDVSLLSI     EN GVFEVQATSGDTHLGGEDFDYKIVRQLIKAFKKKH     GIDVSDN   NKALAKLKREAEKAKRALSSQMS T  r IEIDsFV D G I   d           LSETLTRAKFEELNLDLFKK TLKPVEKVLQDSGLEKKDVDDIVLVGGSTRIP  k VQQLLESYF  d
32 HUMAN   ER   217 E KNILVFDLGGGTFDVSLLTI     DN GVFEVVATNGDTHLGGEDFDQRVMEHFIKLYKKKT     GKDVRKD   NRAVQKLRREVEKAKA LSSQHQ A  r IEIEsFY E G E   d           FSETLTRAKFEELNMDLFRS TMKPVQKVLEDSDLKKSDIDEIVLVGGSTRIP  k IQQLVKEFF  n
33 GIALA        211 A KNILVFDCGGGTHDVSIL S     VDsGVFEVLATAGNTHLGGEDFDRRLLDHFIAIFKKKN   9 G DKAKDmavKKAISRLRREIEAGKRQL STAS S vq IVVD SLID G I  df           SEL VTRAKFEELNIDLFKK SIKPVEQVLRDAKLKTTDIDEVVLVGGSTRIP  k IRQLLQDYF  n
34 GIALA        194 E RNILIFDLGGGTFDVSLLTV  dp SS GVFEVKATAGDTHLGGEDFDSRVVNYFIAEFKKKH     GKDISGS   NRAMRRLRTACEEAKRTLSSSQQ A  s IEIE SLFE G I   d           FFTNITRAKFEDLCIDLFRK CLDPVDRVLRDSKLGKNDVHDIVLVGGSTRIP  k VQQMLRDFF  s
35 TRIVA        190 E RHVLIFDLGGGTFDVSLLDI   d G  GMFEVRATAGDTHLGGEDFDSRLVDYFADRFQKKF     KCDLRES   PRALRTLRTACERAKRTI STAA N 13                            FQDTISRAMFENLNDELFRS TLGPVAQVLKDAKMSKSDVTDIVLVGGSSRIP  r VQQLLQEFF  n
36 PLAFA        204 E KNILIFDLGGGTFDVSLLTI     ED GIFEVKATAGDTHLGGEDFDNRLVNFCVEDFKRKN   r GKDLSKN   SRALRRLRTQCERAKRTLSSSTQaT    IEID SLFE G I   d           YSVTVSRARFEELCIDYFRD TLIPVEKVLKDAMMDKKSVHEVVLVGGSTRIP  k IQTLIKEFF  n
37 ENTHI   CYT  195 E KNVLIFDLGGGTFDVSLLAI     DD GVFEVKASNGDTHLGGEDFDNRLVNHFIAEFKRKY     KKDISGN   ARAVRRLRTACERAKRTLSS AA T an IEVD QLFD G I   d           FYTSITRARFEELNIDLFKS TIGPVERVLQDAKLDKGSIDDVVLIGGSTRIP  k VVQLLQDFF  n
38 MAIZE        196 E KNVLIFDLGGGTFDVSLLTI     EE GIFEVKATAGDTHLGGEDFDNRMVNHFVQEFKRKN     KKDISGN   PRALRRLRTACERAKRTLSSTAQ T  t IEID SLFE G I   d           FTPRSSRARFEELNMDLFRK CMEPVEKCLRDAKMDKSSVHDVVLVGGSTRIP  k VQQL QDFF  n
39 YEAST        188 E EHVLIFDLGGGTFDVSLLSI     ED GIFEVKATAGDTHLGGEDFDNRLVNHFIQEFKRKN     KKDLSTN   QRALRRLRTACERAKRTLSSSAQ T  s VEID SLFE G I   d           FYTSITRARFEELCADLFRS TLDPVEKVLRDAKLDKSQVDEIVLVGGSTRIP  k VQKLVTDYF  n
40 HUMAN        192 E RNVLIFDLGGGTFDVSILTI     ED GIFEVKSTAGDTHLGGEDFDNRMVNHFIAEFKRKH     KKDISEN   KRAVRRLRTACERAKRTLSSSTQ A  s IEID SLYE G I   d           FYTSITRARFEELNADLFRG TLDPVEKALRDAKLDKSQIHDIVLVGGSTRIP  k IQKLLQDFF  n
                     -******1111*1******          ****1 ** 1** 111***1  **  *   ***       * **  *      ***** ***1**1 **1*    *    **** ***         * * *      * - **1**** * - ** * * -*    **** * - ** *****11 *1**      *   *
```

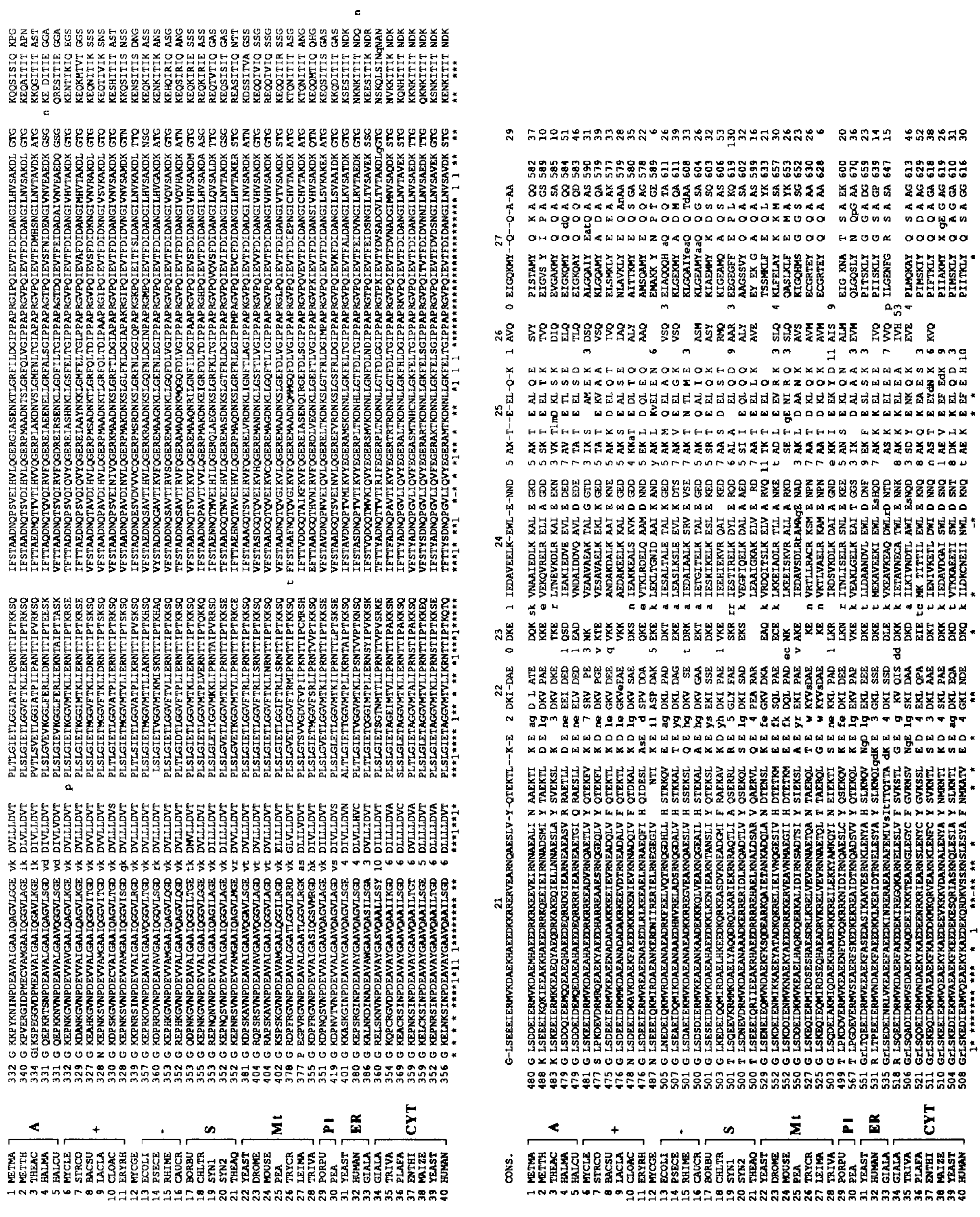

**Fig. 2.** Multiple alignment of 40 HSP70 proteins from prokaryotic and eukaryotic sources. Blocks of alignment are *numbered* and aligned residues shown in *uppercase letters.* Positions completely conserved are indicated by *1.* Positions of conservation index CI ⩾ 0.4 are indicated by an *asterisk.* Positions of negative CI are indicated by a *hyphen.* Unaligned insertions are shown in *lowercase letters* if of length ⩽2, or their length is indicated otherwise. See legend to Fig. 1 for sequence descriptions. Sequence positions are indicated after the sequence names and at the end of the last block of alignment. The last column indicates the length of each carboxy-terminal segment.

human, but much diminished for either with respect to the primitive *Giardia* sequence, 56, 57. Comparing the ER sequences to the eubacterial, chloroplast, and mitochondrial sets gives values at the reduced levels, 41–51.

*Protist SSPA Scores. Trypanosoma, Leishmania* (not shown), *Giardia, Plasmodium, Entamoeba,* and *Trichomonas,* align to the same extent with bacterial and chloroplast sequences in the low range, 37–48 (among the lowest of all SSPA values for the HSP70 proteins at hand). The alignment scores for protists with all CYT-euk sequences rise to the level 63–77, somewhat higher for the vertebrates than for the invertebrates (data not shown).

*SSPA Scores of Cytoplasmic HSP70 from Animals, Plants, and Fungi.* The CYT-euk yeast and human sequences score with CYT-maize at 73 and 77, respectively, and are elevated to 77 for CYT-yeast/CYT-human. Plant HSP70 versus a variety of invertebrates and vertebrates produces SSPA values in the range 70–80 (data not shown). SSPA comparisons of CYT-euk or ER-euk versus Archaea, eubacteria, Pl, and Mt sequences decline to the level 37–51.

*HSP70 repeated sequences.* Among the Cyt-HSP70 sequences in eukaryotes, there are multiple copies often mutually of ⩾80% identity. For example, human and mouse show at least three sequences; bovine—two sequences; *Drosophila melanogaster*—two sequences; *Arabidopsis thaliana*—two sequences; *Plasmodium falciparum*—two sequences; *Leishmania tarentolae*—two sequences; *Trypanosoma brucei*—two sequences. Perhaps surprising, the cyanobacteria SYNP7 (*Synechococcus* sp. and SYNY3 (*Synechocystis* sp.) each contains three HSP70 homologues (designated here P7-1, P7-2, P7-3, and Y3-1, Y3-2, Y3-3). The SSPA score between P7-2 and Y3-2 is 85, while all other pairings align in the reduced range 47–62. Generally, P7-2 and Y3-2 compare to the Mt-HSP70 at the level 50–60, while other homologues have alignments ⩽50. It appears that SYNP7-2 and SYNY3-2 were the original HSP70 in the ancestor of these cyanobacteria. Intriguingly, *E. coli* has two additional sequences that match the DnaK sequence modestly but significantly. These are HSC66 (heat shock cognate) with SSPA score 40 to DnaK and the sequence labeled hypothetical 62 kd protein YBEW with SSPA score 29. *Haemophilus influenzae* also contains two *E. coli* HSP 70 analogs of SSPA scores 81 and 38, respectively. *B. burgdorferi* (BORBU) also possesses two HSP70 sequences with corresponding alignment similarities 62 and 38.

## Multiple Alignment of HSP70 Sequences

The global alignment yields one continuous domain divided into 27 conserved motifs (core blocks) with few unaligned amino acid positions between blocks and few indels (Fig. 2).

*Consensus Sequences of the Multiple Alignment and Conservation Index.* A consensus residue is associated with each aligned position in a core block. The consensus residue is determined as the amino acid which maximizes the (weighted) average similarity score with respect to all residues at that column. In this way, each core block can be summarized by a vector of consensus residues. A conservation index (CI) is computed for each column of the block, reflecting the degree of conservation of the amino acids at the aligned positions. To compute the conservation index the similarity matrix is normalized as by Karlin and Brocchieri (1996), producing the set of normalized scores $s^*(\alpha,\beta) = s(\alpha,\beta)/\sqrt{s(\alpha,\alpha)}\cdot\sqrt{s(\beta,\beta)}$ for amino acids α and β such that identities score 1.0 and substantially dissimilar residues entail a negative score >−1.0. The CI at each position is calculated as the average pairwise normalized similarity value over each column of the *r* aligned sequences: $\text{Cl} = (2/r(r-1)) \sum_{1 \leq i < j \leq r} s^*(i,j)$.

*Primary Signature Segment.* A dramatic result of the alignments is the absence from the Archaea and Gram(+) sequences of core blocks 5 and 6 about 25 amino acids (aa) long. The remaining sequences are significantly aligned at average CI = 0.55 in core block 5 and average CI = 0.37 in core block 6. An additional three to eight residues following block 6 diverge among all the sequences. This result, observed first by Gupta and collaborators (e.g., Gupta and Golding 1993; Gupta et al. 1997), argues that for HSP70, the Archaea, especially the methanococcal and halobacterial sequences, resemble Gram(+) more than other eubacterial sequences. The greater similarity of halobacteria to Gram(+) is also observed in comparing the genome signature among prokaryotic and eukaryotic sequences. Specifically, the halobacterial genome is found to be quite similar to the Gram(+) *Streptomyces griseus* genome and significantly more divergent from all other prokaryotic and eukaryotic genomes (Karlin and Cardon 1994; Karlin and Mrázek 1997b).

*Second Signature Segment.* Core block 13 (4 aa's) is absent from all eukaryotes (ER and CYT forms) but significantly aligned (average CI = 0.64) among all other sequences including the Mt and Pl sequences.

*Conserved Core Blocks.* Inspection of Table 1 reveals that core blocks 2–4, 8–11, 13, 14, and 16–19, contained in the first three-fourths of the protein, are very strongly conserved across all sequences at average CI values exceeding 0.5 (mostly ⩾0.65) and several blocks extend without interruptions to lengths ⩾25 aa, suggesting that these blocks are functionally/structurally important.

*Moderately Conserved Core Blocks.* Blocks 6, 12, 15, 20, and 21, are conserved at the moderate levels CI = 0.37, 0.46, 0.41, 0.42, and 0.44, respectively. The final group of blocks 22–27 register CI values reduced to the level 0.18–0.31 with more variable insertion residues.

**Table 1.** HSP70 consensus sequence[a]

| Bl. | Seq. | Len. | Pos. | Sequence | Ins. | CI |
|---|---|---|---|---|---|---|
| | | | | | 0–67 | |
| 1 | 37 | 3 | 5 | GKV | 0 | 0.35 |
| 2 | 40 | 28 | 8 | IGIDLGTTNSCVAVMEGGKPKVIANAEG | 1 | 0.64 |
| 3 | 40 | 10 | 37 | RTTPSVVAFT | 1 | 0.73 |
| 4 | 40 | 30 | 48 | DGERLVGDPAKRQAVTNPENTIFSVKRLIG | 0 | 0.59 |
| 5 | 27 | 5 | 78 | RRFDD | 2 | 0.55 |
| 6 | 28 | 16 | 85 | VQKDMKHVPYKVVKAD | 5 | 0.37 |
| 7 | 40 | 3 | 106 | VEV | 2 | 0.43 |
| 8 | 40 | 78 | 111 | EGKEYTPEEISAMVLQKMKETAESYLGEKVTNAVITVPAYFNDSQRQATKDAGKIAGLNVLRIINEPTAAALAYGLDK | 3 | 0.72 |
| 9 | 40 | 21 | 192 | DKTILVFDLGGGTFDVSILEI | 0 | 0.75 |
| 10 | 40 | 37 | 213 | GDGVFEVKATNGDTHLGGEDFDNRIVDYLVEEFKKEN | 0 | 0.58 |
| 11 | 40 | 31 | 250 | GIDLSKDKMALQRLREAAEKAKIELSSTTQT | 1 | 0.58 |
| 12 | 39 | 10 | 282 | INLPFITAGA | 1 | 0.46 |
| 13 | 30 | 4 | 293 | GPKH | 0 | 0.64 |
| 14 | 40 | 52 | 297 | LEMTLTRAKFEELTEDLIERTLGPVEQALKDAGLSKSDIDEVILVGGSTRMP | 1 | 0.53 |
| 15 | 39 | 9 | 350 | VQQLVKDFF | 0 | 0.41 |
| 16 | 40 | 28 | 359 | GKEPNKGVNPDEAVAIGAAIQGGVLSGD | 3 | 0.69 |
| 17 | 40 | 8 | 390 | DILLLDVT | 0 | 0.83 |
| 18 | 40 | 29 | 398 | PLSLGIETLGGVMTKLIPRNTTIPTKKSQ | 0 | 0.73 |
| 19 | 40 | 70 | 427 | VFSTAADNQTSVEIKVFQGEREMAKDNKLLGRFDLTGIPPAPRGVPQIEVTFDIDANGILNVSAKDKGTG | 0 | 0.67 |
| 20 | 40 | 11 | 497 | KEQKITITNSS | 0 | 0.42 |
| 21 | 40 | 47 | 508 | GLSEEEIERMVKDAEKHAEEDKKRRERVEARNQAESLVYQTEKTLKE | 2 | 0.44 |
| 22 | 40 | 6 | 557 | DKIDAE | 0 | 0.30 |
| 23 | 39 | 3 | 563 | DKE | 1 | 0.31 |
| 24 | 40 | 16 | 567 | IEDAVEELKEWLENND | 5 | 0.22 |
| 25 | 40 | 8 | 588 | AKTEELQK | 1 | 0.28 |
| 26 | 32 | 3 | 597 | AVQ | 0 | 0.18 |
| 27 | 38 | 12 | 600 | EIGQKMYQQAA | 6–130 | 0.20 |

[a] Core blocks produced in the alignment of the 40 HSP70 proteins. Bl., block number. Seq., the number of sequences included in the block. Len., the length of the block. Pos., the starting position of the block in the consensus sequence. Sequence, the consensus sequence for each block. Ins., the average number of residues between successive blocks. Insertions are represented in the consensus sequence by the average number of *X*'s between blocks. Insertions at the beginning and end of the sequence are arbitrarily set to 4 and 3, respectively. CI, average conservation index of the block.

*Insertions Between Blocks.* Within the aligned parts there are scattered indels. Longer insertions are particularly evident in the two eukaryotic primitive sequences from *Giardia lamblia,* GIALA-ER (9 aa's inserted between blocks 7 and block 8 and between block 10 and block 11; 8 aa's between block 8 and block 9) and GIALA-CYT (10 nonaligned aa's between block 16 and block 17). Most eukaryotic sequences have longer insertions than prokaryotic or organellar sequences between block 7 and block 8 (4–9 aa's vs. 0–4), block 8 and block 9 (3–8 vs. 0–4), and block 16 and block 17 (3–10 vs. 2).

*Amino-Terminal Sequences.* Almost all eubacterial sequences align except for a single N-terminal amino acid (exceptions, *Mycoplasma* MYCGE unaligned for 5 aa's and *Chlamydia* CHLTR unaligned for 6 aa's). At the N terminus all the Mt sequences have a range of 26 to 52 unaligned aa's, putatively enveloping the leader sequence pertinent to translocation to the Mt organelle. The long amino-terminal targeting sequence for Mt is missing from the Mt-like sequence of *T. vaginalis*; PEA-Pl involves 67 unaligned aa's at the N terminus and the ER-euk are unaligned for 49, 27, and 12 aa's. Strikingly, the CYT-euk sequences (except the protist PLAFA) have rather short unaligned peptides.

*Carboxyl-Terminal Variability.* The C-terminal HSP70 sequences manifest variable unaligned lengths, of about 10–50 aa's. However, *Synechococcus* 1 contains a massive 130-aa unaligned terminal sequence.

*Regions and Residues of Highest Conservation.* It is of interest to highlight the most conserved residues (CI $\geqslant$ 0.95) for each core block of the multiple alignment which contain all sequences but at most one.

(i)Twenty conserved glycine positions are prominent including one homotripeptide glycine (in block 9) and two glycine doublets (blocks 10 and 14), all perfectly conserved. A third highly conserved glycine doublet appears in block 18 (CI = 1.00 and 0.81). These glycine diresidues may be critical in establishing structural hinge connections or may facilitate alternative tertiary conformations. Concerning this phenomenon, see the discussion of RecA-like sequences by Brendel et al. (1997) and Brocchieri and Karlin (1998).

(ii) The next most conserved specific residues comprise the aliphatics A (13 positions), V (11), L (10), and I (7). These presumably correspond to hydrophobic core positions.

(iii) Conserved charged residues include D (7 positions), E (3), R (3), and K (2).

(iv) There are no conserved C, H, M, and W amino acids.

Segments of at least three contiguous positions that are highly conserved (CI ⩾ 0.95) are positions 2–7 (GIDLGT) of block 2, five perfectly conserved (see Fig. 2); positions 38–41 (PAYF) of block 8; the succession 67–70 (PTAA) of block 8; positions 10–13 (GGGT) of block 9, emphasizing three contiguous glycines; 16–18 (LGG) of block 10; 44–47 (LVGG) of block 14; 17–19 (GAA) of block 18; and 19–21 (GER) of block 19. The perfect conservation of the basic pair K at position 26 and R (except for one S) at position 27 in block 4 may be functionally important. With the exception of the sequence in blocks 18 and 19, all these conserved segments belong to the ATP-binding domain of HSP70, which extends up to block 16.

## Individual Sequence Similarity to the Consensus Sequence

Within each block individual sequences or sequence groups can be compared to the consensus to assess conservation. For example, eukaryotic sequences score globally with the consensus differently (by at least 0.50) than prokaryotic sequences in the following blocks.

| | Block | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 7 | 11 | 12 | 15 | 23 | 24 |
| Euk. | 1.92 | 2.93 | **3.20** | 2.01 | 0.64 | **3.68** | **3.93** | **2.34** |
| Pro. | **2.89** | **4.05** | 2.26 | **3.65** | **3.55** | 2.36 | 2.13 | 1.75 |

Thus, in the foregoing segments the average similarity assessments of prokaryotic and eukaryotic sequences with the global consensus differ markedly. Generally there are minor fluctuations in average similarity scores among the different subgroups of the prokaryotes.

*Multiple Alignment and SSPA Analysis of Subfamilies.* The multiple alignment protocol was implemented separately for each ‘‘natural’’ group listed next: Archaea (METMA, METTH, THEAC, HALMA, HALCU); Gram(+) (high G+C; MYCLE, STRCO; low G+C; BACSU, LACLA, CLOAC, ERYRH; *Mycoplasma* MYCGE); Gram(−) (γ-type, ECOLI, PSECE; α-type; RHIME, CAUCR); singular (BORBU, CHLTR, SYN1, SYN2, THEAQ); mitochondrial (YEAST, DROME, MOUSE, PEA, TRYCR, LEIMA, TRIVA); chloroplast (PORPU, PEA); eukaryotes—ER isoform (YEAST, HUMAN, GIALA); and eukaryotes—CYT isoform (GIALA, TRIVA, PLAFA, ENTHI, MAIZE, YEAST, HUMAN).

The multiple alignment and consensus sequences of the eight subgroup sets of sequences is given in Fig. 3. Table 2 gives the SSPA values among the consensus sequences of the eight groups. The following comparisons stand out.

(i) The halobacteria, thermoplasma, and methanobacteria generating the Archaea consensus align significantly with the Gram(+) consensus (SSPA value 71).

(ii) The Pl vs. the singular consensuses score dramatically high (74) compared to all other consensus sequence comparisons. This undoubtedly reflects the intrinsic endosymbiotic connections of the chloroplast with cyanobacterial sequences.

(iii) The weakest among all alignment scores, 50, involve the CYT-euk consensus compared to the Archaea, in contrast with other proteins relating Archaea with eukaryotes (e.g., Brown and Doolittle 1997).

(iv) All pairwise group consensus comparisons produce better alignments than do individual pairwise sequence comparisons.

(v) The Gram(+) consensus has its highest alignment score (73) with the singular consensus.

*Multiple Alignment of Consensus Sequences.* The global alignment consensus (Table 1) against the consensus of the consensus sequences (see Fig. 3) gives the impressive SSPA value of 87. The multiple alignment of all eight consensus groups divides into two domains. Domain I consists of five blocks. The signature sequence distinguishing Archaea and Gram(+) described in the global multiple alignment is modified when comparing among the consensuses. Thus, the Gram(+) consensus does not align in blocks 4 and 5 but Archaea diverges only in block 5. Domain II contains 17 core blocks, 6–22. Both eukaryotic consensuses do not align in blocks 9 and 11. The prokaryotic, mitochondrial and chloroplast consensuses do not align in block 12.

## Functional Properties and the Multiple Alignments

The global multiple alignment consensus sequence *S* displayed in Fig. 2 (length, 612 aa, with X substituted for unaligned positions) was investigated by the SAPS (statistical analysis of protein sequences) program (Brendel et al. 1992). The following sequence features stand out.

(i) *S* shows a highly significant mixed charge cluster [for definitions and interpretations, see Karlin 1995)] extending over positions 511–576 from block 21 to block 24, which contains 14 {K or R}, 24 {E or D}, and 25 uncharged residues, and features the hypercharge run EEDKKRRER (at position 526). Eukaryotic sequences, including frog HSP70, *Drosophila* HSP70, human HSP70, *Drosophila* HSP82, yeast HSP90, and chicken

```
                          1                                2                       3                        4                    5
  Cons.           VIGIDLGTTNSCVAVMEGGKPKIIANAEG  1 RTTPSVVAF  2 DGERLVGDPAKRQAVTNPENT  0 FFSVKRLIGRR  0 FNDPEVQKDIKHMPYKV 16

1 Archaea      7 ILGIDLGTTNSAFAVMEGGDPTIIPNAEG xx KTTPSVVAF xd DGEMLVGEPAKRQAVTNPDNT    IYSIKRHMGXX  8
2 Gram(+)      6 VIGIDLGTTNSCVSVMEGGEPKVIANPEG  n RTTPSVVAF  k NGEILVGDAAKRQAVTNP DT 18
3 Gram(-)      5 IIGIDLGTTNSCVAIMDGKTPKVIENAEG  a RTTPSIIAY te DGERLVGQPAKRQAVTNPQNT    LFAIKRLIGRR    FEDPEVQKDKGMMPYKI 14
4 Singular     6 VVGIDLGTTNSCVAVMEGGKPVVIANAEG  t RTTPSVVAF xk NGERLVGQMAKRQAVLNPENT    FYSIKRFIGRR    F D EVTDEIKRVPYKV 16
5 Mt          35 VIGIDLGTTNSCVAVMEGKKPKVIENAEG  a RTTPSVVAF xk DGERLVGMPAKRQAVTNPQNT    FFATKRLIGRR    FDDEEVQKDIKNVPYKI 15
6 Pl          14 VVGIDLGTTNSAIAAMEGGKPTIIPNAEG  x RTTPSVVAY tk NGDKLVGQIAXRQAVINPENT    FYSVKRFIGRK    QN  EISQEXKQTSYNV 16
7 ER          18 VIGIDLGTTYSCVAVMKNGRVEIIANEQG  n RITPSYVAF tx DGERLIGDAAKNQVPANPENT    IFDVKRLIGRK    WNDPSVQKDIKFLPFKV 19
8 CYT          8 AVGIDLGTTYSCVGVWQNDRVDIIANDQG  n RTTPSYVAF  t DTERLIGDAAKNQVAMNPENT    VFDAKRLIGRR    FNDPAVQSDMKHWPFKV 17
                 **1111111*1*******     *1 1 *1    **111**1*     *** 1*1  1**1** 11**1     *  11 *1*        1* *  *-*- ****
                                                                                    Domain I
                                                                       6                                              7
  Cons.          KQFTPEEISAMVLQKMKEDAESYLGEKVTKAVITVPAYFNDSQRQATKDAGTIAGLNVMRIINEPTAAALAYGLDK-K  2 DKTILVFDLGGGTFDVSILEI  0

1 Archaea     89 KEYTPQEISAFILQKIKKDAEAYLGEEIKKAVITVPAYFNDNQRQATKDAGTIAGFDVERIVNEPTAASLAYGLDX E dk DQTILVYDLGGGTFDVSIMDF
2 Gram(+)     84 KDYTPQEISAMILQKLKSDAESYLGEKVTKAVITVPAYFNDAQRQATKDAGKIAGLNVERIINEPTAAALAYGLDK X xx EQKILVFDLGGGTFDVSILEL
3 Gram(-)    109 QXXAPPQVSAEVLQKMKKTAEDYLGEPVTEAVITVPAYFNDAQRQATKDAGRIAGLEVKRIINEPTAAALAYGLDK G ex XRKIAVYDLGGGTFDVSIIEI xx
4 Singular   110 KQFTPEEISAMILRKMKEDASKYLGEKVTEAVITVPAYFNDSQRQATKDAGRIAGLEVKRIINEPTAAALAYGLDK K  g NETILVFDLGGGTFDVSILEI
5 Mt         140 KQYSPSQIGAFVLMKMKETAEAYLGKTVKNAVVTVPAYFNDSQRQATKDAGQIAGLNVLRVINEPTAAALAYGMDKxK    DKIIAVYDLGGGTFDISILEI
6 Pl         118 KDFAPEEISAQVLRKLVDDASKFLGDKVTKAVITVPAYFNDSQRQATKDAGKIAGLDVLRIINEPTAASLAYGFDK K  n NETILVFDLGGGTFDVSILEV
7 ER         127 KVFSPEEISAMVLTKMKEIAEDYLGEKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDK K  4 EKNILVFDCGGGTFDVSLLSI
8 CYT        114 KQFTPEEISSMVLTKMKETAESYLGKKVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDK K  3 ERNVLIFDLGGGTFDVSLLTI
                 *-  1 **** *1 1**  1* *11* ** 11*11111111*111111111 111* 1-1**111111**111*1*          ****1*111111*1** *
                                                                Domain II
                                        8                           9               10                   11
  Cons.          GDGVFEVKATNGDTHLGGDDFDQRIVDYLVDEFKKEHGIDL  1 KDK  0 QALQRLREAAEKAKIELSSTTQTEINL  0 PFITA  2

1 Archaea    189 GGGVFEVKSTNGDTDLGGDDWDHAIIDYLADEFEKEHGIDL  r EDR     QALQRLREAAEKAKIELSTTKETNINL    PFITV td
2 Gram(+)    184 GDGVFEVKATAGDNRLGGDDWDQKIVDYLVDEFKKENGIDL  s KDK     MAMQRLKEAAEKAKIELSSSTETSINL    PFITA xe
3 Gram(-)    211 GDGVFEVKSTNGDTFLGGEDFDMRIIDYLVDEFKKEQGVDL  r NDK     LALQRLKEXAEKAKIELSSSQQTEVNL    PYITA da
4 Singular   209 GDGVFEVKATNGDTHLGGDDFDHKIVNWLAEEFKKEEGIDL  r KDR     QALQRLKEAAEKAKIELSGVSETEINL    PFITA da
5 Mt         239 SNGVFEVKATNGDTHLGGEDFDLALLDYIVSEFKKETGIDL  s KDR     MALQRIREAAEKAKCELSSTMQTEINL    PFITA da
6 Pl         217 GDGVFEVLSTSGDTHLGGDDFDKQIVDWLXXDFKQDEGIDL  x KDK     QALQRLTETAEKAKMELSNLTQTNINL    PFITA tq
7 ER         229 DNGVFEVLATSGDTHLGGEDFDQRIMEHFIKIFKKKNGIDV  8         KAISKLRREVEKAKXALSSQHSTRIEI
8 CYT        215 DXGIFEVKATAGDTHLGGEDFDNRLVNHFIQEFKRKHKKDI  4         RALRRLRTACERAKRTLSSSAQASIEI
                   1*111 *1 11* 111*1*1- ***** -*1*** **1*    *1*     1****** *1*11-*11* -** ***     1*11*
                                                        Domain II (continued)
                 12               13                             14                                         15
  Cons.          DS-F  1 DGPKHLEMTITRAKFEELTEDLIERT  2 PCEKALKDAGLSKSDIDEVILVGGSTRMPKVQELVKDFF  0 GKEPNKGVNPDEAVAIGAAVQGGVLSG  0

1 Archaea    268         DGPKHLEQDITRAKFEELTEDLIERT lg PMEQALEDAGLTKSDIDKVILVGGPTRMPQVQKYVEDFV    GKEPEKNVNPDEAVALGAAIQGGVLSG
2 Gram(+)    263         KGPLHLDMKLTRAKFEELTKDLVERT rq PIRQALKDAGLSVSEIDEVILVGGSTRMPAVQELVKNMT    GKEPNKGVNPDEVVAIGAAIQGGVLTG
3 Gram(-)    290         SGPKHLNIKITRAKLEALVEDLVERT ie PCKVALKDAGLKVSDIDDVILVGGQTRMPKVQEKVKEFF    GKEPHKDVNPDEAVAIGAAVQGGVLQG
4 Singular   288         NGPKHLELTLTRAKFEELAEDLIERT kq PCEQALKDAKLSPSDIDEVILVGGSTRMPAVQQIVKELF    GKEPNKGVNPDEVVAIGAAIQGGVLAG
5 Mt         318         SGPKHLNMKITRSKFESLTQDLIKRT ia PCKKCLKDAGVSLKDINEVILVGGMTRMPKVQETVKELF    GKDPSRGVNPDEAVAIGAAVQGGVLRG
6 Pl         296         DGPKHIEKTITRAKFEELCSDLIDKX xx PVNNALKDAKLEAKDIDEVILVGGSTRIPAIQEMVKKLI    GKDPNXTVNPDEVVAIGAAVQAGVLAG
7 ER         305 DS F  i DGIDFSE TLTRAKFEELNLDLFKKT lk PVEKVLQDSDLKKTDIDEIVLVGGSTRIPKIQQLLQDYF  n GKXXSKGINPDEAVAWGAAVQAGVLSG eq
8 CYT        287 DS1F    EG IDFYTSITRARFEELNIDLFRST ld PVEKVLRDAKMDKSDVHDIVLVGGSTRIPKVQQLLQDFF  n GKEPCKSINPDEAVAYGAAVQAAILSG  4
                 11 1     1* *  - *11***1*1 -11* **     1     1*1* * -*******1111 11*1 *1* **       11 *   *1111*11 111*1***1 1
                                                        Domain II (continued)
                                                                                                   16
  Cons.          DVKDILLLDVTPLSLGIETLGGVMTKLIPRNTTIPTKKSQVFSTAADNQPSVEIHVFQGEREMAKDNKSLGRFELTGIPPAPRGVPQIEVTFDIDANGILNVSAKDKGTGKEQKITITGNSG  0

1 Archaea    362 EVKDIVLLDVTPLSLGIETLGGIFTPLIERNTTIPTRKSQIFTTAADNQTSVQIHVFQGERPIASENKSLGRFILTGIPPAPRGVPQIEVTFDIDANGILNVSAKDKGTGKKQSITIEGGAG
2 Gram(+)    357 DVKDVLLLDVTPLSLGIETMGGVMTKLIERNTTIPTSKSQVFSTAADNQPSVDIHVLQGERPMAADNKSLGRFELTGIPPAPRGVPQIEVTFDIDANGIVNVKAKDLGTGKEQSITISSNSG
3 Gram(-)    384 DVKDVLLLDVTXLSLGIETLGGVFTTLIEKNTTIPTKHSQVFSTADDNQSAVTIRVFQGEREMAADNKMLGQFNLEGIPPAPRGVPQIEVTFDIDANGILHVSAKDKATGKEQKITIKANGG
4 Singular   382 EVKDVLLLDVTPLSLGIETLGGVMTKLIPRNTTIPTRKSEIFSTAADNQTAVEIHVLQGEREMAEDNKSLGRFRLDGIPPAPRGVPQIEVSFDIDANGILHVSAKDKGTGREQKITIQGSST
5 Mt         412 DVTDILLLDVTPLSLGIETLGGVFTRLIPRNTTIPTKKSQXFSTAADGQTQVEIKVFQGEREMASDNKLLGEFTLVGIPPAPRGVPQIEVTFDIDANGIVHVSAKDKATGKEQQITIQSSGG
6 Pl         390 DVKDIILLDVTPLSLGIETLGGVMTKIIPRNTTIPTKKSEVFSTAADGQPNVEINVLQGEREFTKDNKSLGTFRLDGIPPAPRGVPQIEVKFDIDANGILSVKAXDKGTGKEQDITITGAST
7 ER         405 DXXDIVLLDVTPLTLGIETTGGVMTPLIERNTYIPTKKSQIFSTAQDNQPTVKIKVYEGERSMTKDNNLLGNFDLTGIPPAPRGVPQIEVTFEIDANGILRVSATDKGTGKEESITITNDRG  r
8 CYT        389 AVQDLLLLDVTPLSLGIETAGGVMTVLIPRNTTIPTKKSQVFTTYADNQPGVLIQVYEGERAMTKDNNLLGKFELTGIPPAPRGVPQIEVTFDIDANGILNVSAEDKXTGKKNKITITNDKG  r
                 ** 1**11111*1*11111 11**1-*1 *11*111**1* 1*1**1*1  1 1 1 *111 ** *1* 11 1-1 1111111111111*1*111111* 1*1 1* 11*** 111    *
                                                            Domain II (continued)
                                         17                                18                    19               20     21      22
  Cons.          LSDEEIERMVKDAEKHAEEDRKKKERVEVRNQAESLIYQTEKTL  4 DKIDAED-KEKIEKAINEL 12 IEAKTEELEKVVQEI  1 SKMY 10 SGDE  3 EKDE         2

1 Archaea    484 LSDDEIEEMQEEAEQHAEEDRKRKEEIEIRNNAESMIYRAEKTL  4 DKVD ED1KEDVNNEIEDL 12 IKTKTEELSETIQEI  g KQMY 20               601
2 Gram(+)    479 LSDEEIDRMVKDAEKNAEEDRKRKEEVEVRNEAEQLVYQTEKTL  4 XKVPEEE KEKVEKAKDEL 11 IKAKTEKLSQVAQEL  s QKMY 21               596
3 Gram(-)    506 LSDAEIEKMVKDAEANAEADRKKRELVEARNQGDALLHSTEKAL  4 DKLDAGD KTAIESALTAL 12 IEAKTQELAQVSQKL  g EAMY 16               619
4 Singular   504 LSEEEIQRMIKDAEAHAEEDKQRKERIELKNQADSLIYQSEKAL  4 DKISAEE KKRIESHIKEL 11 IKAATEELQKAMYEK 39                      634
5 Mt         534 LSKDEIEKMIKDAEKHAEEDRKKRELIEVRNNAETMINDTEKQL  4 EKLPAEE CKKIKDEISDL 13 IRKATDKLQKASMEL 20           SGGE  3 EGEE 658   kq
6 Pl         512 LPGDDVERMVKEAEXXXXXDKEKREXIDTKNQADSICYQTEKQI  4 DKIXXXI KNKIXXXIGEL 11 IEXXXXXLNNEIMEI 15                      618
7 ER         528 LTEEEIDRMVKEAEKFAEEDKXXKERVESRNELENYAHSLKNQV  7 DKLSSED KETMKEAVKEX 14 FEEKKKKLEKVVQPI  1 SKLY 10 SGDE  3 EKDE 651   1
8 CYT        512 LSKEDIDRMVQDAEKYKAEDEKVKKKVEAKNSLENYCYSVKNTI  5 DKIQAED KEKIEKAVNET 14 FENKQKELEGVCNPI  m SKMY 39               651
                 1* *****1***11--   1* -**-** *1  *    *  ** *     1*-     *  * --*-*     **-    1   -        *1    11-1    1- 1
                                                         Domain II (continued)
```

**Fig. 3.** Alignment of consensus sequences obtained from each group of organisms. See legend to Fig. 2 for symbols and Table 2, footnote a, for composition.

HSP80, show at least one charge cluster (Karlin et al. 1990). Several of the CYT-euk HSP70 sequences possess two distinct charge clusters (see below). In contrast, the *E. coli* HSP70 does not contain a charge cluster of any kind.

(ii) The aggregate frequency of charged {K, R, E, D} residues in *S* is 30.6%, which occurs in the extreme high 5%, i.e., only 5% of all protein sequences (among 50,000 current in SWISS-PROT) have a charge frequency exceeding 30.6%. All but three (HALMA, HALCU, SYN1) sequences have a total charge between 24 and 30% (mostly 26.5–28.5%).

(iii) *S* contains only a single C (cysteine) and a single W (tryptophan) residue.

The Archaea and Gram(+) sequences feature a significant mixed charge cluster in the location of the charge cluster of the consensus sequence. In contrast, the

**Table 2.** SSPA scores of HSP70 group consensuses[a]

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Archaea | — | 71 | 65 | 68 | 64 | 63 | 52 | 50 |
| 2. Gram (+) | | — | 69 | 73 | 66 | 64 | 53 | 57 |
| 3. Gram (−) | | | — | 65 | 71 | 60 | 53 | 51 |
| 4. Singular | | | | — | 68 | 74 | 53 | 55 |
| 5. Mt | | | | | — | 62 | 54 | 51 |
| 6. Pl | | | | | | — | 51 | 53 |
| 7. ER | | | | | | | — | 71 |
| 8. CYT | | | | | | | | — |

[a] SSPA scores for eight consensus sequences of HSP70 subgroups: Archaea is the consensus from five archaeal sequences; Gram(+) from six sequences and *Mycoplasma genitalium;* Gram(−) from four sequences; singular from five unrelated eubacteria; Mt from six mitochondrial sequence and one mitochondrial-like sequence from *T. vaginalis;* Pl from two chloroplast sequences; ER from three endoplasmatic eukaryotic sequences; CYT from seven cytoplasmic eukaryotic sequences.

four Gram(−) proteobacterial sequences of the alignment do not have a charge cluster (even at the 5% significance level). Among the singular bacterial sequences, only CHLTR and THEAQ show the corresponding charge cluster. Also the Mt sequences are variable in this respect. The animal Mt (mouse, *Drosophila*) and *Trichomonas* Mt-like sequence have a concordant charge cluster, whereas the Pea Mt, *Trypanosoma* Mt, and *Leismania* Mt sequences do not possess a charge cluster. Both chloroplast HSP70 sequences contain the corresponding charge cluster. All ER-euk and CYT-euk HSP70 sequences have the analogous charge cluster. Strikingly, most of the CYT-euk sequences carry two separated charge clusters featuring the mixed charge cluster in the usual region and a second charge cluster in human and yeast about positions 244–272.

Molecular chaperones (including HSP70) facilitate the process of protein folding for a wide array of polypeptides. Along these lines, HSP70 proteins are thought to bind with a high affinity to putative hydrophobic patches exposed in unfolded polypeptides, thus preventing aggregation or premature folding. As pointed out above, the major heat shock protein in eukaryotes (HSP70) is distinguished by multiple charge clusters, a highly unusual property of protein sequences that is present in less than 4% of all eukaryotic proteins (Karlin 1995). It is plausible that these charge cluster regions of HSP70 might, to some extent, interact ionically to help orient and position the protein to interact in a hydrophobic manner with special residues of the target protein in its unfolded state. The peptide binding specificity of the ER-euk HSP70 (also known as BiP) seems to be to recognize special hepta- or octapeptides of an improperly folded protein (Flynn et al. 1991). BiP binds to the target peptide via its carboxyl domain. We hypothesize that the substrate binding attributes of BiP explicitly center on the charge cluster. In the target peptide preference for binding is an enrichment of tryptophan residues or, sometimes, other aromatic residues (Blond-Elguindi et al. 1994). Major aromatic residues (W, Y, and F) engage in cationic–aromatic electrostatic interactions where the charge cluster of BiP can be fundamental (Karlin et al. 1994). Moreover, misfolded proteins may also permit charge–backbone interactions of HSP70 with the target peptide. The rarity of W residues (a single position) in the consensus sequence perhaps serves to curtail HSP70 from self-binding with malconformational consequences and avoiding W may also help to maintain HSP70 in a flexible state.

The heat shock proteins including HSP70, HSP90 (human, chicken), HSP108 (chicken, mouse), cognate HSP80 similar to HSP82 (*Drosophila*), HSP80–84 in plants, and HSP82 (yeast) all possess mixed hypercharge runs, supporting the hypothesis of a role of charge in target protein interaction, sorting, transport, and secretion (Karlin 1995).

## Evolutionary Implications

Two measures have dominated analyses of similarities and dissimilarities among various genomes (for a critique see Karlin et al. 1997). (i) In 16S rRNA comparisons, the genes span only about 1500 to 1800 nucleotides, of which less than half are retained in attempting to develop informative alignments. (ii) Protein sequence comparisons also require alignable segments. The amount of sequence available from the ensemble of all proteins is much greater than that of 16S rRNA. The results of such analyses are mixed and often conflicting (e.g., see Gupta and Golding 1996; Doolittle 1996; Karlin et al. 1997). Several authors emphasize lateral gene transfer as a major mechanism for interpreting evolutionary relationships among taxa. This applies, inter alia, to discussions of glutamine synthetase (Brown et al. 1994), glutamate dehydrogenase (Benachenhou-Lahfa et al. 1993), and EF-1α (Baldauf et al. 1996).

Gupta and collaborators rely heavily on HSP70 alignments for comparing taxa; for an alternative perspective see Brown and Doolittle (1997). The HSP70 family displays an amino acid segment of about 25 aa's [called signature sequence by Gupta and Golding (1993) and Gupta et al. (1997)] absent from Archaea and Gram(+) sequences but present and reasonably aligned in eukaryotes and Gram(−) eubacterial sequences. On this basis Gupta and Golding (1993) and Gupta et al. (1997) suggest that Archaea and Gram(+) sequences have a specific relationship whereas Archaea are more divergent from Gram(−) species.

Considering Archaea as a ‘‘coherent’’ group in the consensus sequences Table 2 indicates greater similarity of the archaeal consensus to the Gram(+) consensus than to the Gram(−) consensus. However, the specific SSPA scores indicate that the methanogen METMA HSP70 sequence has a SSPA score relative to eubacterial HSP70 sequences consistently higher than the SSPA scores of the archaeal THEAC, HALMA, and HALCU sequences with the corresponding eubacterial sequences. Notably, METMA alignment scores with Gram(+) (range, 61–70; the highest score obtained with the *Clostridium* sequence CLOAC) are better than scores of METMA with Gram(−) sequences (55–64) and of Gram(+) with Gram(−) sequences (52–64). METMA aligns with the halophile Archaea in the range 55–58. The circumstance of the eocyte (*Sulfolobus*-like) prokaryotes is unresolved since the corresponding HSP70 sequences, if extant, have not been ascertained. These results suggest that the archaeal species are fundamentally heterogeneous. Archaeal sequences tend to be closer to the Gram(+) sequences than to the Gram(−), in line with Gupta and Golding (1993) and Gupta et al. (1997), but the phylogenetic distinctness of Archaea and/or their monophyletic character is not supported by the HSP70-based phylogeny (cf. Gupta and Singh 1994). For example, comparisons of the HSP70 proteins place the *Halobacteria* closer to the *Streptomyces* of high G+C Gram(+) bacteria and *Methanobacteria* closer to *Clostridium* (see also Gupta and Golding 1993; Gupta and Singh 1994; Karlin 1994). Similar results associating archaea and Gram(+) apply to comparisons of the two isoforms of glutamine synthetase GS-I and GS-II (Brown et al. 1994), the isoform of glutamate dehydrogenase GDH-I (Brown and Doolittle 1997), and the FGARAT protein (Gupta and Golding 1996).

The lowest SSPA scores for the archaeal sequences are realized with the eukaryotes, at variance with the hypothesis of Woese et al. (1990) based on 16S rRNA gene comparisons. Actually, in terms of the SSPA values for HSP70 sequences, the halophiles appear significantly eubacterial.

Based on HSP70 sequences Gupta et al. (1997) contend that Gram(−) eubacteria form a phylogenetically coherent group more similar to the eukaryotes than to the Archaea/Gram(+) groups. Contrary to Gupta and Golding (1996) and Gupta et al. (1997), the comparisons to the eukaryotic sequences do not discriminate Gram(+) from Gram(−) sequences. In our analysis, eukaryotic HSP70 sequences seem to be equally far from Gram(−) and Gram(+) according to SSPA scores consistent with the nature of the *second signature* segment (global Block 13) absent in eukaryotes but extant and reasonably aligned in all prokaryote and organelle sequences. Indeed, these blocks to not separate Gram(+), Gram(−) and Archaea which are substantially aligned in the second signature segments, whereas eukaryotic sequences diverge in this region and align separately.

The SSPA analysis of the HSP70 mitochondrial (Mt) sequences separates the protist homologues from the animal–plant–yeast homologues, perhaps suggesting a polyphyletic origin and/or secondary modifications among mitochondrial genomes (cf. Karlin and Mrázek 1997a). Compared to all prokaryotes, the protist Mt sequences score best with α-type proteobacteria, 52 and 58, though significantly lower than nonprotist Mt do, 62–66. Our comparisons are consistent with the hypothesis that the Mt organelle derives as an endosymbiont from an α-type proteobacterium. A caveat: we have provided data [genomic signature comparisons (Karlin and Campbell 1994)] and analysis of phenotypic similarities supporting the hypothesis that the source of animal Mt is more likely a *Sulfolobus*-like prokaryote. As yet, a HSP70 homologue in a *Sulfolobus* species has not been identified.

Three amitochondriate protist clades, trichomonads (e.g., *Trichomonas vaginalis*), diplomonads (e.g., *Giardia lamblia*), and microsporidia (e.g., *Varimorpha necatrix*) are among the earliest-branching lineages of the eukaryotes which primitively either lack mitochondrial organelles or lost mitochondrial function (Germot et al. 1996; Roger et al. 1996). Another deep branch includes *Entamoeba histolytica*. These organisms live under anaerobic conditions. *Trichomonas vaginalis* possesses a hydrogenosome organelle (double membrane, no DNA) but contains an analogue of pyruvate:ferredoxin oxidoreductase that uses pyruvate as a major substrate and oxidizes it to acetyl-CoA, which is converted to acetate + ADP. In this process, hydrogenase specific to the hydrogenosome combines electrons and protons yielding molecular hydrogen. This energy system is classical in *Clostridium* species.

Hydrogenosomes (Hy) are found in rumen-dwelling and free-living ciliates and some fungi. They are known to exist in Trichomonads and probably other amitochondrial protists. The current models of these amitochondrial lineages express the HSP70 and/or HSP60 chaperonins, which align significantly with the corresponding Mt proteins (Germot et al. 1996; Roger et al. 1996, 1998). Views on the origin of the hydrogenosome organelle range from a relic Mt, a higher-order Mt, or a modified Mt. A current hypothesis has Mt derived from an aerobic bacterial source, whereas Hy would be derived from an anaerobic bacterial source, presumably in independent endosymbiotic events. However, a new hypothesis can be proposed indicating that the forebear was a chimeric fusion prokaryote composed from a *Clostridium*-like prokaryote with a hydrogenosome-like capacity and a *Sulfolobus*-like prokaryote. The data and analysis supporting this hypothesis will be elaborated in a future publication.

A single HSP70 homologue was detected in *Entamoeba histolytica* (Ortner et al. 1992), which is most similar to higher eukaryote HSP70 cytoplasmic sequences (SSPA scores, 74–77), with significantly lower similarity scores to other protist (*Giardia lamblia*, *Trichomonas vaginalis*, and *Plasmodium falciparum*) sequences (61–69) and ER sequences (55–61). The ENTHI sequence alignments to Archaea, mitochondria, and chloroplasts show much lower SSPA scores, in the range

40–49. These comparisons suggest that *E. hystolytica* likely never possessed a mitochondrial organelle. *T. vaginalis* encodes two HSP70 sequences, one Mt-like (TRIVA-Mt) (Germot et al. 1996) and the other closest to the eukaryotic cytoplasmic sequences (TRIVA-CYT) (sequence in GenBank). In particular, TRIVA-Mt aligns with the eukaryotic Mt sequences (SSPA scores, 60–63) and at the same level as the α-proteobacterial types RHIME (61) and CAUCR (60). Intermediate similarities are scored with γ-proteobacteria (55, 56), BORBU (55), and other protist mitochondrial sequences (57 with TRYCR-Mt and 55 with LEIMA-Mt). All other alignments have SSPA values ≤53. Similar SSPA score orderings are attained for the Mt-like sequence of the amitochondriate *V. necatrix*. The cytoplasmic TRIVA sequence is closest to the other cytoplasmic sequences (SSPA scores, 65–75), except for GIALA (57). The alignments with the ER sequences are in the range 52–58. All other SSPA scores for TRIVA are significantly lower, 39–47. These results have been interpreted in terms of an original mitochondrion in *T. vaginalis* and *V. necatrix* that was lost (cf. Germot et al. 1996). However, the possibility of horizontal transfer of these Mt-like sequences from an α-proteobacterium cannot be ruled out.

Parenthetically, *T. vaginalis* also encodes two very similar (SSPA value 73) HSP60 sequences, one of which is localized to the hydrogenosome. *G. lamblia* possesses two HSP70 sequences of the ER-isoform and CYT-isoform. These sequences align relatively poorly with all prokaryotic sequences.

The source of Mt organelles have been largely inferred from sequence comparisons of the protein families HSP60 and HSP70 (Gray and Spencer 1996, Viali and Arakaki 1994). However, there are strong suggestions that the Mt functioning HSP60 sequences are not part of any endosymbiont event but a result of lateral transfer probably from an α-proteobacterial progenitor. This conclusion relates to an impressive number of duplicated HSP60 among the classical α-proteobacteria, whereas no duplications among HSP60 sequences are identified from other proteobacterial genomes (data not shown).

The SSPA values for the two chloroplast (Pl) sequences aligned with the classical Gram(−) and Gram(+) and singular eubacteria attain the levels 49–61 (but with SYN2, 73 and 73). These values are consonant with the hypothesis that Pl organelles originate as an endosymbiont from the cyanobacteria.

On the basis of rRNA gene comparisons (Woese et al. 1990), the Archaea are deemed monophyletic and closest to eukaryotes. This conclusion is consistent for certain protein comparisons, e.g., the elongation factor EF-1α and EF-2G families (Iwabe et al. 1989; Creti et al. 1994) and the eukaryotic and archaeal RecA-like sequences of Rad51/Dmc1/RadA (Brendel et al. 1997). However, HSP70 and other protein comparisons challenge the monophyletic character of the Archaea. Results also contesting the pure monophyletic hypothesis from other protein sequence comparisons are given by Benachenhou-Lahfa et al. (1993) in terms of glutamate dehydrogenase sequences and Brown et al. (1994) for glutamine synthetase. Observations of variegated sequence similarities are often interpreted as evidence of lateral gene transfer in the evolutionary history of an organism. Consistent with HSP70 results, genomic signature comparisons (Karlin and Cardon 1994; Karlin and Mrázek 1997b) have the halobacteria closest to the *Streptomyces* sequences, ‘‘moderately similar,’’ and the next twice as distant are *M. tuberculosis* and *M. leprae* sequences. Halobacterial sequences are very distant from the archaebacterial sequences of *Sulfolobus* sp. and *M. thermoautotrophicum.* The classical bacterial genome sequences are generally very distant from *Sulfolobus* sp.

There are many uncertainties concerning divisions among prokaryotes. The HSP70 comparisons of CAUCR (*Caulobacter crescentus*) with RHIME (*Rhizobium meliloti*), both classified as α-type proteobacteria in terms of rRNA genes and HSP70 proteins, are high, 76% identity, whereas the corresponding comparisons for the DnaA protein are low, 30% identity (data not shown). Based on analysis of different protein, gene, or noncoding sequences, different phylogenetic (tree) reconstructions not uncommonly result for the same set of organisms. From this perspective there is a 16S RNA tree, a HSP70 tree, a DnaA tree, etc. The assumption of constant rates of evolution may be violated for different proteins and for different sites within a protein sequence [the problem of unequal rate effects (e.g., Lake 1994; Grishin 1995)]. Moreover, chimeric origins, recombination, transpositions, inversions, lateral transfer, and fusion events may complicate the evolutionary history. There are probably also effects due to ecological, physiological, or other selection forces that putatively engender some convergent evolution that varies from gene to gene. Translation of sequence similarities into evolutionary relatedness and branching order will always be tentative, as the underlying assumptions about mutation rates, selective forces, and gene transfer events are uncertain.

*Acknowledgments.* We are happy to acknowledge helpful comments on the manuscript from B.E. Blaisdell, V. Brendel, and C. Burge. S.K. is supported in part by NIH Grants 5RO1GM10452-33 and 2RO1HG00335-09, NSF Grant DMS9403553-002, and SB-Stanford Research Project 1997.

## References

Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 93:7749–7754

Benachenhou-Lahfa N, Forterre P, Labedan B (1993) Evolution of glutamate dehydrogenase genes: Evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J Mol Evol 36:335–346

Blond-Elguindi S, Cwirla SE, Dower WJ, Lipshutz RJ, Sprang SR, Sambrook JF, Gething MJ (1994) Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. Cell 75:717–728.

Boorstein WR, Ziegelhoffer T, Craig EA (1994) Molecular evolution of the HSP70 multigene family. J Mol Evol 38:1–17

Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89:2002–2006

Brendel V, Brocchieri L, Sandler SJ, Clark AJ, Karlin S (1997) Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. J Mol Evol 44:528–541

Brocchieri L, Karlin S (1998) A symmetric-iterated multiple alignment of protein sequences. J Mol Biol 276:249–264

Brown JR, Doolittle WF (1997) Archaea and the Prokaryote-to-Eukaryote transition. Microbiol Mol Biol Rev 61:456–502

Brown JR, Masuchi Y, Robb FT, Doolittle WF (1994) Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J Mol Evol 38:566–576

Creti R, Ceccarelli E, Bocchetta M, Sanangelantoni AM, Tiboni O, Palm P, Cammarano P (1994) Evolution of translational elongation factor (EF) sequences: Reliability of global phylogenies inferred from EF-1 alpha(Tu) and EF-2(G) proteins. Proc Natl Acad Sci USA 91:3255–3259

Doolittle WF (1996) At the core of the Archaea. Proc Natl Acad Sci USA 93:8797–8799

Flynn GC, Pohl J, Flocco MT, Rothman JE (1991) Peptide-binding specificity of the molecular chaperone BiP. Nature 353:726–730

Germot A, Philippe H, Le Guyader H (1996) Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc Natl Acad Sci USA 93:14614–14617

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

Gray MW, Spencer DE (1996) Organellar evolution in Evolution of Microbial Life eds. Roberts DM, Sharp P, Alderson G, Collins MA, Cambridge Univ Press. 109–126

Grishin NV (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 41:675–679

Gupta RS, Golding GB (1993) Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. J Mol Evol 37:573–582

Gupta RS, Golding GB (1996) The origin of the eukaryotic cell. Trends Biochem Sci 21:166–171

Gupta RS, Singh B (1994) Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr Biol 4:1104–1114

Gupta RS, Bustard K, Falah M, Singh D (1997) Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. J Bacteriol 179:345–357

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Karlin S (1994) Statistical studies of bimolecular sequences: score-based methods. Phil Trans R Soc Lond B 344:391–402

Karlin S (1995) Statistical significance of sequence patterns in proteins. Curr Opin Struct Biol 5:360–371

Karlin S, Brocchieri L (1996) Evolutionary conservation of RecA genes in relation to protein structure and function. J Bacteriol 178: 1881–1894

Karlin S, Campbell AM (1994) Which bacterium is the ancestor of the animal mitochondrial genome? Proc Natl Acad Sci USA 91:12842–12846

Karlin S, Cardon L (1994) Computational DNA sequence analysis. Annu Rev Microbiol 48:619–654

Karlin S, Mrázek J (1997a) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227–10232

Karlin S, Mrázek J (1997b) Prokaryotic genome-wide comparisons and evolutionary implications. In: de Bruijn FJ, Lupski J, Weinstock G (eds). Bacterial genomes: Physical structure and analysis, in press. Chapman, Hall, New York

Karlin S, Blaisdell BE, Brendel V (1990) Identification of significant sequence patterns in proteins. In: Doolittle R (ed) Methods of enzymology, Academic Press, San Diego, CA

Karlin S, Weinstock G, Brendel V (1995) Bacterial classifications derived from RecA protein sequence comparisons. J Bacteriol 177: 6881–6893

Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899–3913

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc Natl Acad Sci 91:1455–1459

Ortner S, Plaimauer B, Binder M, Wiedermann G, Scheiner O, Duchene M (1992) Humoral immune response against a 70-kilodalton heat shock protein of *Entamoeba histolytica* in a group of patients with invasive amoebiases. Mol Biochem Parasitol 54:175–184

Roger AJ, Clark CG, Doolittle WF (1996) A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis.* Proc Natl Acad Sci USA 93:14618–14622

Roger AJ, Svard SG, Tovar J, Clark CG, Smith MW, Gillin FD, Sogin ML (1998) A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc Natl Acad Sci USA 95:229–234

Viale AM, Arakaki AK (1994) The chaperonin connection to the origins of the eukaryotic organelles. FEBS LETT 341:146–151

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eukarya. Proc Natl Acad Sci USA 87:4576–4579