Heat shock protein 60 sequence comparisons: Duplications, lateral transfer, and mitochondrial evolution

Samuel Karlin* and Luciano Brocchieri

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

Contributed by Samuel Karlin, August 11, 2000

Heat shock proteins 60 (GroEL) are highly expressed essential proteins in eubacterial genomes and in eukaryotic organelles. These chaperone proteins have been advanced as propitious marker sequences for tracing the evolution of mitochondrial (Mt) genomes. Similarities among HSP60 sequences based on SIGNIFICANT SEGMENT PAIR ALIGNMENT calculations are used to deduce associations of sequences taking into account GroEL functional/structural domain differences and to relate HSP60 duplications pervasive in α-proteobacterial lineages to the dynamics of lateral transfer and plasmid integration. Multiple alignments with consensuses are determined for 10 natural groups. The group consensuses sharpen the similarity contrasts among individual sequences. In particular, the Mt group matches best with the classical α-proteobacteria and closely with *Rickettsia* but significantly worse with the rickettsial groups *Ehrlichia* and *Orientia*. However, across broad protein sequence comparisons, there appears to be no consistent prokaryote whose protein sequences align best with animal Mt genomes. There are plausible scenarios indicating that the nuclear-encoded HSP60 (and HSP70) sequences functioning in Mt are results of lateral transfer and are probably derived from an α-proteobacterium. This hypothesis relates to the plethora of duplicated HSP60 sequences among the classical α-proteobacteria contrasted with no duplications of HSP60 among other clades of proteobacterial genomes. Evolutionary relations are confounded by differential selection pressures, convergence, variable mutational rates, site variability, and lateral gene transfer.

HSP60 | GroEL | mitochondria | protein similarity | gene duplication

Heat shock protein 60 (GroEL) is an abundant essential protein in all *Escherichia coli* life stages (e.g., during log growth and in stationary phase) and in most bacteria (1). The family of HSP60 proteins is well studied for its role as chaperone facilitators of protein folding and in rescuing the cell from stress conditions (e.g., see review in ref. 2). HSP60 proteins are ubiquitous in eubacterial cells and also function in the mitochondrial (Mt) and plastid organelles of eukaryotes. In particular, HSP60 (and HSP70) facilitate bidirectional traffic between the mitochondrion and the cytoplasm. HSP60 protein structures form heptameric rings that dimerize in a barrel-like complex with a central plane of symmetry (3, 4). Each monomer folds into a structure divided into three domains: Domain E occupies the Equatorial section of the double ring, Domain A consists of the Apical part, and the Intermediate Domain I connects the previous two. Each Domain contributes a specific function in the HSP60 complex. Domain E includes most of the connections between monomers of the same ring and between rings and contains the ATP/ADP/Mg^{2+}-binding pocket. Domain I closes on the binding pocket, providing essential residues for ATP hydrolysis. Domain A binds to HSP10 (GroES) and to the target substrate. Proteins of the extended HSP60 family (eubacterial GroEL, eukaryotic Tcp1, and archaeal thermosome) are highly expressed. This property, observed in many evolutionary lineages, may be related to the mobility of the HSP60 genes within and between genomes. Duplication and horizontal gene transfer of HSP60 may promote functional adaptation and differentiation.

Several chaperone and degradation proteins that function in Mt organelles of eukaryotes have been proposed as good marker sequences for tracing the evolution of Mt genomes (5, 6). The current Mt endosymbiont hypothesis, inferred from rRNA gene sequence comparisons, proposes that Mt genomes were acquired from a Gram-negative α-proteobacterium. Viale and Arakaki (7), by using the neighbor-joining algorithm (8) applied to HSP60 protein sequences, proposed that Mt sequences are most related to the *Rickettsia tsutsugamushi* sequence, now reclassified as *Orientia tsutsugamushi* (9). Andersson *et al.* (10) proposed *Rickettsia prowazekii* (RICPR in the SwissProt genus–species nomenclature; see Fig. 1) as the likely endosymbiont forebear of the Mt organelle. Recently, this interpretation has been supplanted by some unspecified member of the α-proteobacterial clade (e.g., refs. 11, 12). However, analysis of different protein families reveals no consistent prokaryotic organism most similar to mitochondria (see below). Increasing writings now advocate the view that the first eukaryotes and mitochondria arose in unison (13–15). It is also recognized that genomes of many organisms, especially prokaryotes and primitive eukaryotes, consist of ''heterogeneous unions,'' ''consortia,'' and chimeras to which lateral transfer and/or close associations have substantially contributed (13–16).

The main objective of this paper is to evaluate similarities between HSP60 sequences on the basis of SIGNIFICANT SEGMENT PAIR ALIGNMENT (SSPA) calculations, with special attention to Mt evolution (the SSPA method is formally described in *Methods*), taking into account GroEL structural properties, paralogs, and influences of lateral transfer.

Methods

SSPA. For convenience, we outline the SSPA protocol (for elaborations, see refs. 17, 18). A pairwise amino acid similarity matrix s(*i,j*) (e.g., BLOSUM62; for review, see ref. 19) is used to score pairwise amino acid similarities. Given two sequences to be aligned, pairs of sequence segments are identified that attain an aggregate score exceeding the score attained for corresponding random sequences of the same composition with probability $<$ 0.01. Extant high-scoring matching segments among protein sequences putatively imply conservation because of essential biological structure/function. The global similarity between two protein sequences is scored as follows: first, all HSSPs (high-scoring segment pairs) significant at the 1% level are identified. Next, the HSSPs are combined into a consistent alignment. The

Abbreviations: Mt, mitochondrial; SSPA, SIGNIFICANT SEGMENT PAIR ALIGNMENT.

*To whom reprint requests should be addressed. E-mail: karlin@math.stanford.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

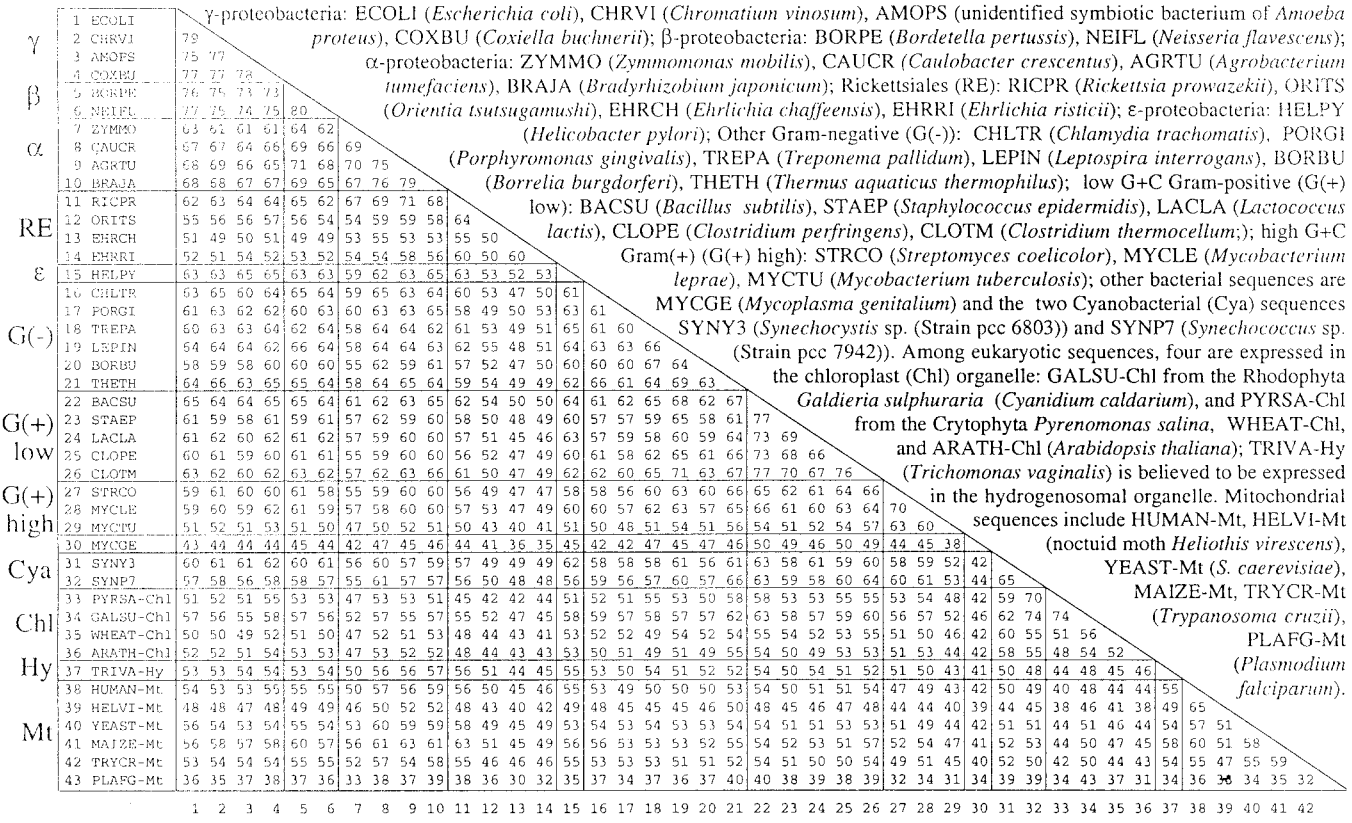

Fig. 1. SSPA similarities of bacterial and organellar HSP60 sequences.

alignment (SSPA) score is the maximal value with respect to all sets of consistent matching sequence segments calculated by summing HSSP segment scores and then normalizing to allow comparison of proteins of different sizes and quality. For the sequence pairings with at least one hit (i.e., a segment having a significantly high SSPA match), additional segments are identified by using a lower probability threshold (typically 0.50). The use of the second reduced threshold helps to fill in regions between the more significant HSSP. The SSPA scores are used to deduce groupings of sequences. A group is deemed coherent if the SSPA scores within the group almost invariably exceed the SSPA scores with sequences not in the group, and if the scores with sequences of other groups are consistent for all members of the groups.

Data. From more than 150 distinct HSP60 sequences found, a culled collection was formed retaining 43 representative sequences of mutual SSPA scores not exceeding 80% (Fig. 1). These include sequences from six β- or γ-proteobacteria, four α-proteobacteria, four *Rickettsia/Orientia/Ehrlichia*, one ε-type (*Helicobacter pylori*), six singular Gram(−), five low G + C Gram(+), three high G + C Gram(+), one mycoplasma, two cyanobacteria, seven sequences classified as Mt-like, including one hydrogenosomal (Hy) sequence from the aMt eukaryote *Trichomonas vaginalis* (TRIVA), and four Plastid sequences.

The classical α-proteobacterial sequences divide into two major subgroups, one important in nitrogen fixation (e.g., *Rhizobium* spp.) and a second found predominantly in soil and marine habitats and performing anoxygenic photosynthesis (e.g., *Rhodobacter* spp.). A tentative third group, the Rickettsiales, including the obligate intracellular parasites *Rickettsia*, *Orientia*, and *Ehrlichia* genera, has been grouped with α-proteobacteria, apparently on the basis of 16S rRNA gene comparisons. However, genome signature and protein comparisons (see Figs. 1 and 2) indicate drastic discrepancies between classical α and Rickettsiales (15). The classical α genomes are pervasively of high G + C content (>60%), whereas rickettsial genomes are of low G + C content (<32%).

Results and Discussion

SSPA Values Among HSP60 Bacterial Sequences. We implemented the SSPA methodology to ascertain similarities among the 43 representative HSP60 sequences. Regions of divergence and conservation among sequences were analyzed through their multiple alignment (secured with the INTERALIGN protocol) reported in ref. 20.

HSP60 sequences of the β and γ-proteobacterial sequences (ECOLI, CHRVI, AMOPS, BORPE, COXBU, and NEIFL; see Fig. 1) have SSPA scores mutually high in the range 73–80 and <71 (usually much less) with the other sequences. Classical α-proteobacterial sequences (CAUCR, AGRTU, and BRAJA) also cluster (SSPA values 75–79). ZYMMO vs. classical α produces similarity scores 67–70 but significantly lower (61–64) with (β + γ)-proteobacteria, relegating ZYMMO to be a distant relative of the α-proteobacteria. The alignment of the RICPR sequence relative to classical α-proteobacterial HSP60 sequences produces SSPA scores about 68 to 71. However, contrary to RICPR comparisons, the three sequences of ORITS, EHRCH, and EHRRI align to the α-sequences at the diminished SSPA levels 53–59. HSP60 sequence similarities unambiguously separate *Orientia* and *Ehrlichia* from each other (SSPA value 50) and from all currently available bacterial groups.

The Gram(+) sequences divide into two groups: low G + C (BACSU, STAEP, LACLA, CLOPE, and CLOTM) and high G + C (STRCO, MYCLE, but excluding MYCTU), with SSPA scores 66–77 within groups. The SSPA scores between these

Mt:	Hum	Gga	Dme	Nem	Fun	Pla	Cre	Ccr	Kpl	COXI[1]
RICPR	62	61	62	54	56-64	70,71	64	70	38	BRAJA/Mt >> RICPR/Mt ≈ Other α/Mt
α	58-69	56-66	58-68	51-59	52-70	65-70	57-64	62-70	34-38	α/α (62-92) ≥ α/RICPR (62-68)

Mt:	Hum	Gga	Dme	Nem	Fun	Pla	Ccr	Lta	COXII[2]
RICPR	45	48	47	36,37	40-43	43,45	47	23	RICPR/Mt >> α/Mt
α	31,37	38,40	37,38	26-29	29-34	32-37	38,40	19,22	α/α (53) >> α/RICPR (33,35)

Mt:	Mam	Gga	Dme	Nem	Fun	Zma	Pwi	Ccr	Aca	COXIII[3]
RICPR	49,50	49	48	38,40	39,41	44	52	47	47	α/Mt ≥ RICPR/Mt
PARDE	49,53	54	54	39,40	38-46	51	56	51	45	

Mt:	Pla	Pwi	Ram	NADH DEHYDROGENASE UNIT 2[4]
RICPR	24,26	26	28	α/Mt >> RICPR/Mt
RHOCA	34,37	37	38	α/α (72) >> α/RICPR (25,26)
PARDE	32,35	33	34	

Rat	Ama	Mpo	Ram	NADH DEHYDROGENASE UNIT 4
24	38	47	48	α/Mt >> RICPR/Mt
28	43	51	54	α/α (80) >> α/RICPR (45,55)
27	44	50	53	

Mt:	Hum	Ncr	Mpo	Ram	NADH DEHYDROGENASE UNIT 5
RICPR	29	33	40	43	α/Mt ≈ RICPR/Mt
RHOCA	29	29	40	42	α/α (67) >> α/RICPR (40,42)
PARDE	30	31	40	44	

Hum	Cel	Pla	Ram	NADH DEHYDROGENASE UNIT 7[5] - RICPR/Mt>>α/Mt
69	64	45,72	76	except with ARATH.
60	58	59,64	66	α/α (85) >> α/RICPR (66)
60	57	59,64	66	

Mt:	Hum	Ncr	Ram	NADH DEHYDROGENASE UNIT 11
RICPR	47	47	50	α/Mt ≈ RICPR/Mt
RHOCA	47	46	49	α/α (77) >> α/RICPR (43,44)
PARDE	46	47	48	

Mt:	Mmu	Sce	Mpo	Ram	ATPASE - F1[6]
RICPR	66	63	68	65	α/Mt >> RICPR/Mt
α	72	70	75	72	α/α (73-86)[9] >> α/RICPR (63-69)

Mt:	Mmu	Fun	Cma	CYTOCHROME C[7]
RICPR	36	36,39	34	α/Mt >> RICPR/Mt
BRAJA	43	42,45	36	

Mt:	Ver	Dme	Fun	Ath	Cre	Ccr	Pfa	CYTOCHROME B[8]
RICPR	52-54	52	48,55	58	50	62	42	α/Mt ≈ RICPR/Mt
α	49-54	48-54	41-56	51-59	44-50	58-64	33-37	α/α (54-86) ≈ α/RICPR (58-70)

	Rpr1	Rpr2	Abr	Rsp	Rme	Bsu	GLUTAMYL-tRNA SYNTHETASE
RICPR1	-	24	45	23	22	30	BACSU/Mt ≥ α/Mt
RICPR2	24	-	26	45	28	36	≥ RICPR/Mt
YEAST-Mt	22	26	24	24	29	31	

	Rpr	α	Sau	Smu	Cab	σ70[9]
RECAM-Mt	6	6	12	11	11	Gram(+)/Mt > RICPR/Mt
RICPR	-	49,50	41	51	51	≈ α/Mt
α	49,50	69	39,43	52	56	α/α (69) >> α/RICPR (49,50)

	Eco	Rca	Rpr	Bca	SUPEROXIDE DISMUTASE (SOD)
RHOCA	53	-	40	45	BACCA/Mt >> α/Mt
RICPR	56	40	-	42	≈ RICPR/Mt
MOUSE-Mt	39	33	33	53	

	Rpr	γ	Gr(-)	Gr(+)	RNA POLYMERASE CHAIN β'[10]
RICPR	-	57-60	48-53	51-53	RICPR/Mt ≥ Gram(+)/Mt
RECAM-Mt	32	28-29	26-29	29-31	

Fig. 2. SSPA Similarities between eubacterial and mitochondrial sequences are indicated for different protein families. Abr, *Azospirillum brasilense*; Aca, *Acantamoeba castellanii*; Ama, *Allomyces macrogynus* (Chytridiomycota); Ath, *Arabidopsis thaliana*; Bca, *Bacillus coldotenax*; Bsu, *Bacillus subtilis*; Cab, *Clostridium acetobutylicum*; Ccr, *Chondrus crispus* (Rhodophyta); Cel, *Caenorhabditis elegans*; Cma, *Cucurbita maxima* (pumpkin); Cre, *Chlamydomonas reinhardtii*; Dme, *Drosophila melanogaster*; Eco, *Escherichia coli*; Gga, *Gallus gallus*; Hum, human; Lta, *Leishmania tarentolae*; Mmu, mouse; Mpo, *Marchantia polymorpha* (liverwort); Ncr, *Neurospora crassa*; PARDE, *Paracoccus denitrificans*, Pfa, *Plasmodium falciparum*; Pwi, *Prototheca wickerhamii* (Chlorophyta); Ram, *Reclinomonas americana*; Rca, RHOCA, *Rhodobacter capsulatus*; Rme, *Rhizobium meliloti*; Rpr, *Rickettsia prowazekii*; Rsp, *Rhodobacter sphaeroides*; Sau, *Streptomyces aureofaciens*; Sce, *Saccharomyces cerevisiae*; Smu, *Streptococcus mutans*; Zma, maize. [1]α, *Bradyrhizobium japonicum*, *Rhodobacter leguminosarum*, *R. sphaeroides*; Nem, nematodes *Ascaris suum* (pig roundworm) and *C. elegans*; Fun, fungi *Emericella nidulans*, *N. crassa*, *S. pombe*, and *S. cerevisiae*; Pla, plants *A. thaliana* and maize; Kpl, kinetoplastida *L. terentolae* and *Trypanosoma brucei brucei*. [2]α, *P. denitrificans* and *R. sphaeroides*; Nem, *A. suum* and *C. elegans*; Fun, *N. crassa*, *S. cerevisiae*, and *S. pombe*; Pla, plants *D. carota* (carrot) and maize. [3]Mam, human and mouse; Nem, *A. suum* and *C. elegans*; Fun, *N. crassa*, *S. pombe* and *Schizophyllum commune*. [4]Pla, *A. thaliana* and liverwort. [5]Pla, *A. thaliana* and wheat. [6]α, *Rhodospirillum rubrum* and *R. capsulatus*. [7]Fun, *Candida glabrata* and *S. cerevisiae*. [8]α, *R. sphaeroides*, *R. capsulatus*, and *R. rubrum*; Ver, vertebrates human, mouse, and chick; Fun, *N. crassa* and *S. cerevisiae*. [9]α, *R. meliloti* and *C. crescentus*. [10]γ, *H. influenzae*, *E. coli*, and *Pseudomonas putida*; Gr(−), Gram(−) BORBU, CHLTR, HELPY, TREPA, and SYNY3; Gr(+), Gram(+) BACSU and *S. aureofaciens*.

groups are in the range 60–66, not dissimilar from their scores with most other bacterial sequences. Other singular Gram(−) bacterial sequences (TREPA, LEPIN, BORBU, THETH, CHLTR, and PORGI) and the two Cyanobacteria SYNY3 and SYNP7 score among themselves in the range 60–67 as they score with most other bacterial sequences including proteobacteria and Gram(+) in support of their classification as isolated and/or early branching bacterial lineages.

Mt HSP60 Sequences. Apart from the outlier sequence PLAFG-Mt from the protist *Plasmodium falciparum*, Mt sequences mutually score in the range 47–65, the highest score being between the two animal sequences HUMAN-Mt and HELVI-Mt (nocturnal moth *Heliothis virescens*), and the lowest score being between HELVI-Mt and TRYCR-Mt. Comparisons of RICPR to Mt HSP60 sequences yield SSPA scores in the range 48–63, similar to the scores of classical α-proteobacteria vs. Mt (50–63). The HSP60 ORITS aligns to Mt sequences about 10 points lower (43–51).

Similarities and Differences in the Structural Domains. As noted earlier, each GroEL monomer possesses three structural Domains: Apical, Intermediate, and Equatorial. Similarity comparisons of SSPA scores were ascertained separately for the sequences over each of the three structural domains of HSP60 (for detailed data, see http://gea.stanford.edu/luciano/hsp60.sspa). The RICPR sequence similarities to Mt sequences parallel those of classical α-proteobacteria in Domains E and A but are lower in Domain I (SSPA values 62–68 vs. 64–73). For ORITS, similarity to Mt is about 10 points lower than that of α-proteobacteria in all three structural domains. ORITS has scores about equal to RICPR in Domain I but divergent in Domains A and E. *Ehrlichia* sequences, in comparisons with Mt sequences, are equivalent to *Orientia* in Domains A and I but are sharply divergent in Domain E, with assessments of similarities 20 points lower than for classical α-proteobacteria and lower than for any other group. The reduced similarities of HELVI-Mt with most other sequences can be attributed to its

Table 1. SSPA values of HSP60 group consensuses

	1	2	3	4	5	6	7	8	9	10
6 $\beta + \gamma$ proteobacteria	100	76	66	59	73	65	63	64	63	64
4 α-proteobacteria	76	100	75	63	74	67	62	64	64	66
RICPR	66	75	100	66	62	58	67	60	58	63
ORITS + 2 *Ehrlichia*'s	59	63	66	100	61	58	51	54	55	55
7 singular Gram(−)	73	74	62	61	100	73	67	70	67	65
5 low G + C Gram(+)	65	67	58	58	73	100	69	68	68	60
3 high G + C Gram(+)	63	62	67	51	67	69	100	65	62	55
2 cyanobacteria	64	64	60	54	70	68	65	100	73	56
4 Chl sequences	63	64	58	55	67	68	62	73	100	59
5 Mt + 1 Hy sequence	64	66	63	55	65	60	55	56	59	100

SSPA similarities are between consensus sequences derived from the alignment of the following groups: (*i*) $\beta + \gamma$ proteobacteria ECOLI, CHRVI, AMOPS, COXBU, BORPE, NEIFL; (*ii*) α-proteobacteria ZYMMO, CAUCR, AGRTU, BRAJA; (*iii*) RICPR; (*iv*) Divergent Rickettsiales ORITS, EHRCH, EHRRI; (*v*) singular Gram(−) CHLTR, PORGI, TREPA, LEPIN, BORBU, THETH; (*vi*) low G + C Gram(+) BACSU, STAEP, LACLA, CLOPE, CLOTM; (*vii*) high G + C Gram(+) STRCO, MYCLE, MYCTU; (*viii*) cyanobacteria SYNY3, SYNP7; (*ix*) chloroplast sequences PYRSA-Chl, GALSU-Chl, WHEAT-Chl, ARATH-Chl; (*x*) Mt and hydrogenosomal sequences HUMAN-Mt, HELVI-Mt, YEAST-Mt, MAIZE-Mt, TRYCR-Mt, TRIVA-Hy. See legend to Fig. 1 for full names.

pronounced divergence within Domain A. The diminished matching exhibited by the ZYMMO sequence vs. classical α-proteobacterial sequences can be explained by differences in Domain E, where ZYMMO vs. Mt scores are 41–48, individually lower than those of ($\beta + \gamma$)-proteobacteria, 45–54, and of α-proteobacteria, 49–60. Within Domain A, the ZYMMO sequence scores like a classical α-proteobacterium (ZYMMO/α-proteobacteria, 80–84). The high G + C Gram(+) *Mycobacterium tuberculosis* is lower in both Domains A and E but not in Domain I**.** *Mycoplasma genitalium* diverges in all three domains. PLAFG-Mt is the most deviant sequence in all three domains, with lowest SSPA scores in Domain E, generally less than 30 and marginally elevated SSPA scores in Domains A and I (30–40).

Multiple Alignment of HSP60 Consensus Group Sequences. We used the multiple alignment program ITERALIGN (20) to identify blocks of alignment among HSP60 sequences (see also http://gea.stanford.edu/luciano/hsp60.alignment). A consensus sequence is derived that best summarizes the residue composition of the alignment. All HSP60 sequences align with few indels. Actually, bacterial sequences match from the N terminus on (apart from one to three residues). By contrast, organelle sequences generally include an expanded N-terminal segment (presumably a peptide leader sequence) of variable length of 23–68 amino acids. The C-terminal region is unaligned or poorly aligned and generally contains repetitive elements (20).

Multiple alignments were determined separately for each of the following HSP60 groups: (*i*) six ($\beta + \gamma$)-proteobacteria; (*ii*) four classical α-proteobacteria; (*iii*) the single RICPR sequence; (*iv*) the "sister" Rickettsial sequences ORITS, EHRCH, and EHRRI; (*v*) seven singular Gram(−); (*vi*) five low C + G Gram(+); (*vii*) three high G + C Gram(+); (*viii*) two cyanobacteria; (*ix*) four Chl sequences; (*x*) five Mt (excluding PLAFG-Mt) + 1 Hy sequence. The sequence names are given in Table 1. The consensus from the multiple alignments of the consensuses produced the impressively high similarity 91% to the global individual consensus. Predictably (18), SSPA values among group consensuses sharpen the contrasts among individual and group sequences (Table 1). In particular, the Mt group aligns best with the classical α-proteobacteria (SSPA score 66) but registers very close scores (64, 65) to other Gram(−) group sequences. The consensus of the group ORITS, EHRCH, and EHRRI aligns best with RICPR (66) but is significantly lower with the consensus Mt sequence (55). Consistent with the endosymbiont hypothesis, the chloroplast sequences score about 73 with cyanobacterial sequences and <68 (typically much less) in comparisons with all other sequences.

Duplications of the HSP60 Gene. Many species contain multiple copies of HSP60, *a priori* paralogs, with high mutual SSPA values. Notably, several α-proteobacterial genomes feature multiple HSP60 copies. Specifically, *Rhizobium meliloti* (RHIME) possesses at least five distinct HSP60 sequences (21, 22) of mutual SSPA scores in the range 75–95; *Bradyrhizobium japonicum* (BRAJA) contains at least five distinct HSP60 sequences (23) with high SSPA scores; and *Rhodobacter sphaeroides* (RHOSH) contains two HSP60 sequences of about 75% identity. Multiple HSP60 sequences also exist in the Cyanobacterium *Synechocystis* sp (two copies), in the Gram(+) *Streptomyces lividus* (two), *M. leprae* (two), and *M. tuberculosis* (two), the respective pairs being all about 80% similar. Strikingly, of the aggregate proteobacterial collection, multiple HSP60 sequences have to date been identified only among α-proteobacteria.

The primitive aMt eukaryote *T. vaginalis* contains two very similar HSP60 sequences of mutual SSPA score 73. One of the proteins functions in the hydrogenosome organelle (24), and the other is of unknown localization. To date, only a single copy of HSP60 has been found in *Giardia lamblia* (25) and in *Entamoeba histolytica* (26) with low similarities to eubacterial and higher eukaryote organellar sequences, in the range 30–35% and 40–45% identity, respectively. *P. falciparum* carries at least two HSP60 sequences (27, 28). One is Mt like, and the other shows highest similarity (42%) to the GALSO (red alga) HSP60 sequence.

A version of HSP60 binds to the rubisco protein in the chloroplast. Multiple such sequences have been to date identified in *Arabidopsis thaliana*, in the pea plant, in *Brassica*, and in the green alga *Chlamydomonas reinhardii*. The Tcp1 and thermosome proteins are recognized as the eukaryotic and archaeal homologues of HSP60 although their SSPA scores with respect to eubacterial HSP60 range only from 0 to 10% identity. In surveying the current complete genomes, we found that *A. fulgidus*, *A. pernix*, and *M. thermoautotrophicum* contain two thermosome sequences, whereas *M. jannaschii* contains a single sequence. *Sulfolobus acidocaldarius* possesses at least two homologues.

Expression of GroEL Genes. The pattern of expression of GroEL genes has been well studied in BRAJA (23). BRAJA GroEL-3 (the third copy, 58 kDa) synthesis is coregulated with the nitrogen fixation system, whose genes are transcribed from a σ^{54} promoter. It is possible that this GroEL protein is a requirement for NifA folding and nitrogenase assembly (29). GroEL-2 and GroEL-4 apparently are constitutively expressed. GroEL-1 is under heat shock control. GroEL-5 expression is regulated independently of NifA by cellular oxygen conditions. Moreover,

GroEL-2, GroEL-4, and GroEL-5 can functionally replace each other (23). Thus, the five GroEL/GroES complexes are differentially regulated, allowing a flexible response to varying environmental conditions and physiological needs. In RHIME, the DNA-binding activity of NodD requires the product of GroELc (22). This gene is encoded on one of the two megaplasmids of RHIME, possibly allowing for lateral transfer during plasmid movements.

Why Multiple Copies of a Gene? (*i*) Gene duplication conceivably can increase the expression level of the encoded protein at various times and places and under special conditions. (*ii*) The duplicated copies can functionally diverge or participate in heterooligomer complexes. This appears to be the nature of some HSP60 structures, for example, HSP60 rubisco-binding proteins in plastids. Duplicated genes freed from functional constraints can evolve faster and adapt to new needs. (*iii*) Duplication may provide insurance against extreme fluctuation of expression and against mutation or other detrimental events. (*iv*) The genome may be simply large enough to tolerate duplicated benign genes. Mechanisms for duplication and mobility include transposition, recombination, conjugation, transformation, and transduction.

It is accepted that DNA sequences can be laterally transferred between organisms (30, 31) and have been transferred in evolution from cytoplasmic organelles to the nucleus and/or between organelles (32). The presence of multiple copies suggests mobility of HSP60 genes in α-proteobacteria. The multiplicity of HSP60 sequences attests to its dynamic character and may suggest high intrinsic potential for lateral transfer from some α-proteobacterium.

Other Protein Sequence Comparisons. It is useful to summarize SSPA similarity values for various classes of protein sequences, emphasizing classical α- proteobacteria, RICPR, and Mt sequences. A manifest conclusion emerging from the data is the lack of a prokaryotic group that is consistently most similar to animal Mt sequences (see Fig. 2).

Proteins encoded in animal Mt genomes. For cytochrome oxidase I (CoxI), CoxIII, ATPase F_1, cytochrome *c*, and NADH units 2 and 4, the Mt sequences match better with at least some classical α-sequences than with RICPR. For the proteins NADH 5, NADH 11, and cytochrome *b* sequences, similarity attainments of α-proteobacteria vs. Mt are about the same as for RICPR vs. Mt. The proteins CoxII and NADH 7 show the alignment inequality RICPR vs. Mt $>$ α vs. Mt.

Mt aminoacyl-tRNA synthetases. Arginyl: yeast Mt vs. γ-proteobacterial sequences reach 19–22% identity, 3-fold better than the comparisons of yeast Mt vs. RICPR showing only about 7% identity. Aspartyl: yeast Mt vs. BORBU attains the SSPA level 31, which dominates yeast Mt vs. RICPR 22. Threonyl: fungal Mt sequences from *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, and *Candida albicans* compared with γ-proteobacteria and BACSU carry 30–36% identity but in alignment to RICPR, 27–29% identity. Tyrosyl: BACSU matches with yeast Mt at 38% identity, whereas RICPR vs. yeast Mt have 28% identity (data not shown). Glutamyl: yeast Mt vs. BACSU carries 31% identity compared with Mt vs. RICPR of 22% identity.

Chaperones and proteases functioning in the Mt (data not shown). For the Lon degradosome gene: (BACSU vs. Mt) 38–40% $>$ (α vs. Mt) 34–38% and RICPR vs. Mt 33–35% identity. For the metallo protease FtsH: Mt sequences tentatively match best with *Streptococcus pneumoniae* with substantially diminished identity to RICPR. ClpP: RICPR vs. γ, 70–75%, RICPR vs. α, 62–66%, indicating for this degradation protein that RICPR is more similar to γ-types than to α-types.

Other proteins. For the proteins DnaA, elongation factor EF-Tu (data not shown), superoxide dismutase (SOD), and the RNA polymerase unit β', RICPR matches to γ bacterial sequences significantly better than in matching to classical α-types. For example, with respect to the detoxification protein SOD, RICPR vs. γ-types align at 50–56% identity, but RICPR vs. RHOCA shows only 40% identity. The σ 70 factor aligns RECAM-Mt poorly with RICPR and α-proteobacteria.

Functional Specialization and Selective Convergence of HSP60 Proteins. Although organisms of recent common origin are expected to exhibit higher sequence similarity, evolutionary relations can be obscured by convergence, lateral gene transfer, variable mutation rates, site variability, etc. Moreover, proteins may be subject to variable selective pressures, depending on physiological and/or ecological conditions. This can cause sequence divergence at different rates in different organisms (the problem of unequal evolutionary rates). A few characters suggest some form of functional differentiation among HSP60 proteins: (*i*) Many HSP60 sequences include the iterated C-terminal tripeptide GGM. The function of these tails is unknown, but similar iterations are present as C-terminal elements in the unrelated HSP70 chaperone proteins. Some HSP60 proteins, however, do not have these repeats but incorporate instead C-terminal tails emphasizing multiple histidines. Notably, MYCTU (*Mycobacterium tuberculosis*) has two HSP60 genes, one coding for a protein with C-terminal GGM elements and a second with multiple histidines. It is possible that the switch between these types corresponds to different functional specializations. (*ii*) Among the ATP/ADP-binding positions of HSP60, position 52 is a conserved lysine (K) in proteobacteria, some nonproteobacterial Gram-negative and Mt sequences, but is a pervasive asparagine (N) in Gram-positive sequences and Cyanobacteria. Many other positions surrounding the ATP/ADP-binding pocket switch from highly conserved residues to other residues in some sequences. Switch positions may suggest different mechanisms of coupling ATP hydrolysis to the substrate refolding process. (*iii*) There are differences in the ways the GroEL protein complex assembles and functions. For example, in mitochondria, cpn60 can function as a single ring, whereas two rings are needed in *E. coli* (33).

Convergent evolution has never been proved between molecular sequences (e.g., ref. 34). However, the same environmental conditions that may suggest the evolution of Rickettsiales into mitochondria may also suggest convergent evolution of the HSP60 and other sequences. The endosymbiotic lifestyle of mitochondria and parasitic lifestyle of Rickettsia correspond to specializations of their metabolism that result in reduction and simplification of their genomic and protein content. Consider also switch positions with respect to ATP-binding sites (20) that may relate to convergent function specialization.

HSP60s of mitochondria and of many Rickettsiales appear to have been subjected to fast evolutionary rates, as reflected by the fact that they generate long branches in phylogenetic tree reconstructions (e.g., ref. 7). At the same time, the estimated common ancestors of these groups are separated by much shorter distances. Similarities involving Mt sequences and Rickettsial sequences may be sufficiently low to place them in the region where phylogenetic information is largely lost. In clustering Mt with Rickettsiales, the phenomenon of long branch attraction may be at play.

Perspectives. Methods have been developed (e.g., refs. 8, 35) that seek to reconstruct evolutionary relations by building tree-like topologies, where branch lengths are putatively proportional to evolutionary time, and the branch topology reflects events of speciation. However, tree-making procedures rest on uncertain assumptions and problematic approximations (36). Results are often influenced by many factors, including problems with alignments, definition of homology, lateral gene transfer, gene

loss, data set, clustering method, and models of evolution. In particular, different protocols have been proposed to estimate evolutionary distances between pairs of sequences based on their sequence similarity.

Our analysis indicates that HSP60 contained in mitochondria is closest (by SSPA Analyses) to the classical α-proteobacteria and RICPR. Does this imply that the Mt is the remnant of an ancient α-proteobacterial organism? This inference is not supported by other genomic characters [e.g., genomic signature (15), other protein comparisons; see text]. Current studies of molecular evolution emphasize lateral gene transfer as a major evolutionary mechanism (14, 30, 31). Lateral transfer among all organisms (bacteria, fungi, plants, animals, protists, etc.) may be involved in promotion of new species (30). The ubiquitous role of lateral transfer in affecting and shaping prokaryotic and eukaryotic species is increasingly appreciated. For example, reacquisition and dissemination of antibiotic resistance genes via mobile genetic elements (e.g., conjugative plasmids, phages, free DNA, and transposons) is an established paradigm of lateral transfer.

Primitive organisms probably engaged in much reduction, acquisition, and lateral transfer of DNA, producing chimeric genomes. We propose that the nuclear encoded HSP60 sequences functioning in Mt are a result of lateral transfer and are probably derived from a classical α-proteobacterial progenitor. One of the requirements for the successful acquisition of a laterally transferred gene is its utility to the recipient organism. HSP60 proteins facilitate folding of a great variety of proteins, acting on substrates whose selection seems to be solely constrained by their size (37). One would expect that genes highly advantageous to the recipient organism bear potential for successful interspecies gene flow. From this perspective, the nonspecificity of HSP60 proteins to their targets makes them likely viable gene acquisitions. In this context, α-proteobacteria possess multiple features that suggest that they may be likely donors of HSP60 genes: (*i*) they possess unusual facility for HSP60 gene duplication and transposition, as indicated by the plethora of HSP60 sequence duplications among classical α-proteobacteria, in contrast to no paralogs of HSP60 sequences in the other clades (γ, β, δ, ε) of proteobacterial genomes; (*ii*) α-proteobacteria establish close spatial and functional relations, often endosymbiotic, with plant eukaryotic organisms. Lateral gene transfer provides a means of quick response to strong selection pressures reflected by virulence factors ranging from toxin production to immune evasions. Examples include Ti plasmids of *Agrobacterium tumefaciens*, nodulation plasmids of *Rhizobium*, and virulence plasmids of *Shigella* and *Yersinia*; (*iii*) copies of α-proteobacterial HSP60 genes have been found to reside on extrachromosomal elements of the α-proteobacterial genome, e.g., the megaplasmids pSyma and pSymb of *R. meliloti* (22).

We thank B. E. Blaisdell and A. M. Campbell for critical reading of the manuscript. This work was supported in part by National Institutes of Health Grants 5R01GM10452–36 and 5R01HG00335–12 and by National Science Foundation Grant DMS9704552–002.

1. Karlin, S. & Mrázek, J. (2000) *J. Bacteriol.* **182,** 5238–5250.
2. Sigler, P., Zhaohui, X., Rye, H. S., Burston, S. G., Fenton, W. A. & Horwich, A. L. (1998) *Annu. Rev. Biochem.* **67,** 581–608.
3. Boisvert, D. C., Wang, J., Otwinowski, Z., Horwich, A. L. & Sigler, P. B. (1996) *Nat. Struct. Biol.* **3,** 170–177.
4. Xu, Z., Horwich, A. L. & Sigler, P. B. (1997) *Nature (London)* **388,** 741–750.
5. Budin, K. & Philippe, H. (1998) *Mol. Biol. Evol.* **15,** 943–956.
6. Gupta, R. S. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 1435–1491.
7. Viale, A. M. & Arakaki, A. K. (1994) *FEBS Lett.* **341,** 146–151.
8. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
9. Tamura, A., Ohashi, N., Urakami, H. & Mijamura, S. (1995) *Int. J. Syst. Bacteriol.* **45,** 589–591.
10. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & Kurland, C. G. (1998) *Nature (London)* **396,** 133–140.
11. Gray, M. W., Burger, G. & Franz Lang, B. (1999) *Science* **283,** 1476–1481.
12. Andersson, S. G. & Kurland, C. G. (1999) *Curr. Opin. Microbiol.* **2,** 535–541.
13. Martin, W. & Müller, M. (1998) *Nature (London)* **392,** 37–41.
14. Lopez-Garcia, P. & Moreira, D. (1999) *Trends Biochem. Sci.* **24,** 88–93.
15. Karlin, S., Brocchieri, L., Mrázek, J., Campbell, A. M. & Spormann, A. M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9190–9195.
16. Moreira, D. & Lopez-Garcia, P. (1998) *J. Mol. Evol.* **47,** 517–530.
17. Karlin, S., Weinstock, G. M. & Brendel, V. (1995) *J. Bacteriol.* **177,** 6881–6893.
18. Brocchieri, L. & Karlin, S. (1998) *J. Mol. Biol.* **276,** 249–264.
19. Brendel, V. (1996) *Adv. Comput. Biol.*, **2,** 121–160.
20. Brocchieri, L. & Karlin, S. (2000) *Protein Sci.* **9,** 476–486.
21. Rusanganwa, E. & Gupta, R. S. (1993) *Gene* **126,** 67–75.
22. Ogawa, J. & Long, R. (1995) *Genes Dev.* **9,** 714–729.
23. Fischer, H. M., Babst, M., Kaspar, T., Acuña, G., Arigoni, F. & Hennecke, H. (1993) *EMBO J.* **12,** 2901–2912.
24. Bui, E. T. N., Bradley, P. J. & Johnson, P. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9651–9656.
25. Roger, A. J., Svard, S. G., Tovar, J., Clark, C. G., Smith, M. W., Gillin, F. D. & Sogin, M. L. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 229–234.
26. Clark, C. G. & Roger, A. J. (1995) *Proc. Natl. Acad. Sci. USA* **95,** 6518–6521.
27. Holloway, S. P., Min, W. & Inselburg, J. W. (1994) *Mol. Biochem. Parasitol.* **64,** 25–32.
28. Syin, C. & Goldman, N. D. (1996) *Mol. Biochem. Parasitol.* **79,** 13–19.
29. Govezensky, D., Greener, T., Segal, G. & Zamir, A. (1991) *J. Bacteriol.* **173,** 6339–6346.
30. de la Cruz, I. & Davies, I. (2000) *Trends Microbiol.* **8,** 128–133.
31. Campbell, A. M. (2000) *Theor. Popul. Biol.* **57,** 71–77.
32. Martin, W. & Herrmann, R. G. (1998) *Plant Physiol.* **118,** 9–17.
33. Nielsen, K. L. & Cowan, N. J. (1998) *Mol. Cell* **2,** 93–99.
34. Doolittle, R. F. (1994) *Trends Biochem. Sci.* **19,** 15–18.
35. Fitch, W. M. (1997) *Syst. Zool.* **20,** 406–416.
36. Brocchieri, L. (2000) *Theor. Popul. Biol.,* in press.
37. Houry, W. A., Frischman, D., Eckerskorn, C., Lottspelch, F. & Hartl, F. U. (1999) *Nature (London)* **402,** 147–154.