Conservation among HSP60 sequences in relation to structure, function, and evolution

LUCIANO BROCCHERI AND SAMUEL KARLIN

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

(RECEIVED September 23, 1999; FINAL REVISION December 10, 1999; ACCEPTED December 10, 1999)

Abstract

The chaperonin HSP60 (GroEL) proteins are essential in eubacterial genomes and in eukaryotic organelles. Functional regions inferred from mutation studies and the *Escherichia coli* GroEL 3D crystal complexes are evaluated in a multiple alignment across 43 diverse HSP60 sequences, centering on ATP/ADP and Mg^{2+} binding sites, on residues interacting with substrate, on GroES contact positions, on interface regions between monomers and domains, and on residues important in allosteric conformational changes. The most evolutionary conserved residues relate to the ATP/ADP and Mg^{2+} binding sites. Hydrophobic residues that contribute in substrate binding are also significantly conserved. A large number of charged residues line the central cavity of the GroEL–GroES complex in the substrate-releasing conformation. These span statistically significant intra- and inter-monomer three-dimensional (3D) charge clusters that are highly conserved among sequences and presumably play an important role interacting with the substrate. Unaligned short segments between blocks of alignment are generally exposed at the outside wall of the Anfinsen cage complex. The multiple alignment reveals regions of divergence common to specific evolutionary groups. For example, rickettsial sequences diverge in the ATP/ADP binding domain and Gram-positive sequences diverge in the allosteric transition domain. The evolutionary information of the multiple alignment proffers attractive sites for mutational studies.

Keywords: evolution; function; GroEL; HSP60; structure

The 60 kD heat shock proteins are ubiquitous abundant proteins of eubacterial genomes. Eukaryotic versions are nuclear encoded but function in organelles: mitochondria (Mt), chloroplasts (Chl), and hydrogenosomes (Hy). The TCP1 protein family of the eukaryotic cytoplasm and the archaeal thermosomes are viewed as HSP60 homologues, albeit sequence similarity between TCP1 and bacterial HSP60 is weak. HSP60 proteins have been vigorously studied for their chaperone function in protein folding and for their cooperation with other chaperones (particularly HSP70) in cellular trafficking (Hartl, 1996). A plethora of HSP60 sequence duplications is observed among classical α -proteobacteria in contrast with no duplications of HSP60 sequences detected in the β , γ , δ , ϵ clades of proteobacterial genomes. Multiple HSP60 sequences also exist in *Cyanobacteria* and in several Gram-positive genomes. More than 150 homologues of HSP60 sequences are currently available with pairwise similarity extending from 40 to 100% at the amino acid level. Our analysis is based on a set of 43 representative sequences of which no two entail more than 80% similarity. For a listing, see the caption to Figure 1.

Several crystal structures of GroEL are available. Two of them show GroEL in its 14 unit complex, one associated with ATP γ S

(Boisvert et al., 1996) and a second in an asymmetrical conformation bound to GroES and ADP (Xu et al., 1997). Each GroEL monomer distinguishes three structural Domains: A (Apical), I (Intermediate), and E (Equatorial). Seven identical GroEL monomers are assembled in rings that dimerize in a 14-meric structure with an equatorial interface formed by the E Domains and two noncommunicating central cavities. Each Domain E also provides a binding pocket for an ATP/ADP molecule and a Mg^{2+} ion. Domain I bridges Domain E to Domain A and Domain A connects to GroES. When GroES attaches to Domain A with ADP present the 7 unit ring attains a modified structural conformation (*cis* conformation) characterized by an expanded central cavity compared to the free monomers (*trans* conformation). On the basis of biochemical and mutational studies (e.g., Fenton et al., 1994), several key function and structure regions have been proposed including residues for binding and hydrolyzing ATP, positions binding the substrate and exposed to the central cavity, monomer intra- and inter-unit contact sites, and connections with GroES. For reviews on data and models on the functioning of the GroEL–GroES complex in interaction with the substrate, see, e.g., Horowitz (1998), Netzer and Hartl (1998), Sigler et al. (1998). The multiple alignment of the 43 representative HSP60 protein sequences coupled to analysis of conservation and variation provides an opportunity to integrate sequence (evolutionary relationships), structure, function, and mutation information.

Reprint requests to: Samuel Karlin, Department of Mathematics, Stanford University, Stanford, California 94305-2125; e-mail: fd.zgg@forsythe.stanford.edu.

Fig. 1. Alignment of HSP60 proteins. Consensus sequence from the multiple alignment of 43 HSP60 sequences aligned with the *E. coli* HSP60 sequence. Lower case refers to residues in the *E. coli* sequence unaligned; numbers in the Consensus sequence indicate the average length of insertions between blocks of alignment. Positions in the structural Domain E are set in brown; Domain I are set in orange; Domain A are set in green. Highly conserved peptides are shown in blue. The conservation of each alignment position is measured by the CI (see Materials and methods). A minus sign (–) refers to CI < 0.0; 4 signifies 0.4 ≤ CI < 0.7; 7 signifies 0.7 ≤ CI < 0.9; 9 signifies 0.9 ≤ CI < 1.0; 1 is CI = 1.0. Good conservation requires CI > 0.4. CI < 0.0 indicates poor conservation. Functional sites: ‡ refers to ATP/ADP and Mg²⁺ binding sites; G refers to GroEL binding positions; R refers to inter-ring contacts; S refers to substrate binding positions (Fenton et al., 1994); + or – refers to cationic and anionic residues, respectively, exposed to the central cavity in the *cis* conformation. Sequences considered: γ -proteobacteria: ECOLI (*Escherichia coli*), CHRVI (*Chromatium vinosum*), AMOPS (bacterium of *Amoeba proteus*), COXBU (*Coxiella burnetii*); β -proteobacteria: BORPE (*Bordetella pertussis*), NEIFL (*Neisseria flavescens*); α -proteobacteria: ZYMMO (*Zymomonas mobilis*), CAUCR (*Caulobacter crescentus*), AGRTU (*Agrobacterium tumefaciens*), BRAJA (*Bradyrhizobium japonicum*); Rickettsiales: RICPR (*Rickettsia prowazekii*), ORITS (*Orientia tsutsugamushi*), EHRCH (*Ehrlichia chaffeensis*), EHRRI (*Ehrlichia risticii*); ϵ -proteobacteria: HELPY (*Helicobacter pylori*). Other Gram-negative: CHLTR (*Chlamydia trachomatis*), PORGI (*Porphyromonas gingivalis*), TREPA (*Treponema pallidum*), LEPIN (*Leptospira interrogans*), BORBU (*Borrelia burgdorferi*), THETH (*Thermus thermophilus*); low G+C Gram-positive: BACSU (*Bacillus subtilis*), STAEP (*Staphylococcus epidermidis*), LACLA (*Lactococcus lactis*), CLOPE (*Clostridium perfringens*), CLOTM (*Clostridium thermocellum*); high G+C Gram(+): STRCO (*Streptomyces coelicolor*), MYCLE (*Mycobacterium leprae*), MYCTU (*Mycobacterium tuberculosis*); other bacterial sequences are MYCGE (*Mycoplasma genitalium*), Cyanobacteria SYN3 (*Synechocystis* sp.) and SYN7 (*Synechococcus* sp.). Eukaryotic sequences, Chl organelle: GALSU-Chl (*Galdieria sulphuraria*), PYRSA-Chl (*Pyrenomonas salina*) WHEAT-Chl, ARATH-Chl (*Arabidopsis thaliana*); TRIVA-Hy (*Trichomonas vaginalis*) is believed to be expressed in the hydrogenosomal organelle; Mt sequences include HUMAN-Mt, HELVI-Mt (*Heliothis virescens*), YEAST-Mt (*Saccharomyces cerevisiae*), MAIZE-Mt, TRYCR-Mt (*Trypanosoma cruzii*), and PLAFG-Mt (*Plasmodium falciparum*).

Results

The results focus on two main perspectives: (1) Analysis and interpretation of the HSP60 multiple sequence alignment. (2) Evolutionary sequence conservation in relation to functional properties of the HSP60 proteins.

Multiple alignment of HSP60 sequences

The complete multiple alignment of the HSP60 sequences is available online at: <http://gea.stanford.edu/luciano/hsp60.html>. Moreover, tables with complete lists of position variability and contacts are available from the Electronic Edition. From the mul-

multiple alignment, a Consensus Sequence is derived which “optimally” summarizes the alignment. Conservation is evaluated for each position by the Conservation Index (CI) and for each sequence conservation is determined by comparison to the Consensus Sequence (see Materials and methods and Brocchieri & Karlin, 1998). The consensus sequence is displayed in Figure 1. Henceforth, all positions will be referred to the *Escherichia coli* sequence coordinates.

Terminal unaligned segments

Eubacterial sequences match from the N-terminus on (except 1 to 3 residues). Organelle sequences commonly include an expanded N-terminal segment (putatively a peptide leader sequence) of variable length in the range 23–68aa's. The C-terminal region is unaligned or poorly aligned and contains GGM (one letter amino acid code) iterations (see Table 1). However, some HSP60 sequences show no vestige of GGM, including Chl sequences, the Mt sequences HELVI-Mt (SWISS-PROT sequence code) and PLAFG-Mt (featuring instead a long acidic tail), MYCGE, and the high G+C Gram(+) sequences STRCO and MYCTU. Intriguingly, GGM iterations are also observed in HSP70 sequences.

Core alignment blocks

Blocks 5–24 constitute the core of the alignment (Table 2). These blocks possess strong conservation and are generally separated by short indels. Blocks 5–8, 23, 24, and the end of Block 22 correspond to Domain E, which includes most inter-monomer and ATP/ADP and Mg^{2+} contact sites. The “Rickettsial-like” sequences ORITS and especially EHRCH and EHRRI are sharply deviant in Block 5 but maintain the ATP/ADP binding residues. We also observe some divergence in all Chl sequences. Block 8, the longest of all core blocks (92aa), is significantly conserved and contains several positions binding ATP/ADP and ligands of the Mg^{2+} ion. The alignment degenerates at the end of Block 8 in the transition from Domain E to Domain I. Block 9, including part of Domain I, is largely conserved in α -proteobacteria and Rickettsiales, but divergent among most Gram(+), Chl, and Cyanobacteria. Block 12 encompasses the conserved transition region from Domain I to Domain A with conspicuous glycines and substrate binding sites occupied by conserved aromatic residues. Blocks 15 to 19 are abundant in conserved substrate binding and exposed charged residues, and Block 16 is the most conserved of all blocks. Blocks 20–21 are poorly conserved among Chl, Mt, and Rickettsiales.

Domain A ends in Block 22, which continues through Domain I into Domain E. Conservation in this block is excellent among proteobacteria. The *Orientia* and *Ehrlichia* sequences register low sequence conservation in the subsequent Blocks 23 and 24. Blocks 18–24 are the most variable among HSP60 sequences.

Unaligned regions between blocks

The four primary unaligned segments, 139–142, 336–340, 430–434, and 470–474, are all exposed at the outside surface of the rings and are probably not essential to the functions of the protein complex.

Highly conserved residues

There are 186 highly conserved residues ($CI \geq 0.75$) in the complete alignment of HSP60 sequences distributed among amino acids as follows.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
$0.75 \leq CI < 0.9$	13	–	3	6	1	10	–	12	5	7	1	2	3	–	5	2	6	13	–	2
$0.9 \leq CI < 1.0$	9	–	3	6	1	6	–	9	6	6	–	2	2	–	2	2	3	7	–	1
$CI = 1.0$	–	–	3	–	–	17	–	–	–	3	–	–	2	–	1	–	2	2	–	–

Among these positions, glycine (G) is outstanding, with a total of 33 conserved positions, including 17 perfectly conserved. Glycine also shows one perfectly conserved triplet (position 414) and one perfectly conserved doublet (positions 374). The next most

Table 1. C-terminal sequences^a

ECOLI	LPKNDADLGAAGGMMGMMGMM
CHRV	EPKDDAPAMPGGMGDMGMMGMM
AMOPS	LPKKDEGMAGAGDMGMMGMMGMMGMM
COXBU	APKKKEESMPGGGDMGMMGMMGMMGMM
BORPE	LMENKPAAAPAMPGGMGMMGMMDF
NEIFL	IPERKPAMPDMGMMGMMGMM
ZYMMO	LPEDKPAPAMPGGMGMMGMMGMMDF
CAUCR	PQEGRRRPAGGGMPGMMGDMDF
AGRTU	LPKKESAMPQMPGGMGMMDF
BRJA	VPKNTGAGGMPGGMGMMGMMGMMDF
RICPR	EPDKEEPMMPMRGGMGMMGMMGMMDF
ORITS	HEEDNNTGNRSGGGVGGGHHGGMGMMDF
EHRCH	VPSKDDANAGAGGGMGMMGMMGGF
EHRRI	LPTKDNGSPSMPGGMGMMGMMGGF
HELPY	IKEEKATPAMPDMGMMGMMGMMGMM
CHLTR	IPEEKPAAPAMPAGMDY
PORGI	KKEDNPAAPAMPGGMGMMGMM
TREPA	IGQKSSSTPPAPDMGMMGMMY
LEPIN	KPDKDAPNPMAGMGGGMGMMGMM
BORBU	IKEEKNTSGGGGYPMDPGMM
THETH	KPEKKESTPASAGAGDMDF
BACSU	KPEENGAGAMPDMGMMGMMGMM
STAEP	IPEPENNEQPGMGMM
LACLA	KPEPAAPAMPMDPSMGMMI
CLOPE	IPEKEMPQGAGMGMDGY
CLOTM	IPEKETSGGPGGAGMGMMY
STRCO	KKEEEPAAGGSHSHGSH
MYCLE	KPEKTAAPASDPTGGMGMMDF
MYCTU	KPAKAEDHDHGHGHAH
MYCGE	VKERKDNSFSE
SYNY3	KPEPAAPAMPDMGMMGMMGMMGMM
SYNP7	KPEPKEKAPAGAGGGMGDFDY
PYRSA-Chl	KKKTCSELSIYWLVLTKYFCLDLLRLYFFLKD
GALSU-Chl	EMNRNMEVRK
WHEAT-Chl	KPKPKPKVAEPAEGQLSV
ARATH-Chl	IKEPEPVPVGNPMDNSGLGY
TRIVA-Hy	VPEEKPAAPAMPMPGMMY
HUMAN-Mt	IPKEEKDPGMDGAMGGMGMMGMMF
HELVI-Mt	MPKQKDPPEYAPIGGGDY
YEAST-Mt	APEPPAAAGAGGMPGMPGMM
MAIZE-Mt	IPKEEAPAPAMGGMGMMDY
TRYCR-Mt	FPKEETPAAGGMMGMMGMMGMDY
PLAFG-Mt	HEKKDKGILDSSINSPNYLSKHRRYKHLHDDE DTDEDEDEDEDEDEDDLDLDDDDYDEDEDEDE DEDEDEDEDEDEDSMNDEYNDE
PLAFA-Mt ^b	FKDSKNEESSQHNSVNSMGDMGMMY
MYCTU1 ^b	KPEKEKASVPGGDMGMMDF

^aAll C-terminal sequences are shown starting from position 524 of the *E. coli* sequence.

^bThese sequences are not included in our 43 sequence data set.

Table 2. Sequence conservation levels in core blocks

	Block																			
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Length	34	6	4	92	11	4	22	48	4	5	35	18	5	5	3	27	17	72	35	58
CI ^a	0.53	0.70	0.68	0.56	0.68	0.57	0.66	0.56	0.63	0.57	0.67	0.80	0.59	0.66	0.47	0.50	0.45	0.69	0.49	0.51
Seq. Cons. ^b	3.26	3.67	4.50	3.38	3.45	3.25	3.64	3.44	3.50	2.80	3.57	4.44	3.80	3.40	3.33	3.04	2.76	3.71	2.97	3.02
ECOLI	3.97	4.17	5.50	3.93	3.55	2.75	4.00	3.44	3.75	3.20	3.69	4.56	4.00	3.80	3.00	3.44	2.82	3.96	2.57	3.43
CHRV1	3.53	4.67	5.50	4.03	3.55	2.25	4.32	3.62	4.25	3.00	4.06	4.83	2.80	3.80	3.33	2.93	3.41	4.07	2.91	3.28
AMOPS	3.35	4.17	5.50	3.67	3.55	2.25	4.18	3.71	2.00	3.20	4.06	4.83	4.00	3.80	1.67	3.41	3.06	4.11	3.31	3.74
COXBU	3.35	4.17	5.50	3.84	3.55	3.00	4.18	3.94	3.50	3.40	3.83	4.78	4.40	3.80	3.33	3.19	3.06	4.26	2.77	3.84
BORPE	3.26	4.17	5.50	3.96	3.73	3.00	4.32	3.67	4.25	4.40	3.63	5.00	4.00	3.80	2.67	3.37	3.12	4.07	2.49	3.53
NEIFL	3.44	3.67	4.50	3.89	3.73	2.50	4.59	3.65	4.25	4.40	3.54	4.83	3.80	3.80	3.33	3.26	2.71	4.12	2.80	3.33
ZYMMO	3.65	n.a. ^c	1.25	2.51	3.45	3.50	4.09	3.85	4.00	2.60	4.20	5.00	3.00	4.00	4.33	4.07	2.82	4.12	3.20	2.83
CAUCR	3.44	4.33	4.00	3.74	4.09	4.75	3.55	3.83	2.50	2.40	4.14	5.00	4.00	4.40	4.33	4.22	3.53	4.06	3.71	2.86
AGRTU	3.79	3.83	5.50	3.78	4.00	2.25	3.59	4.02	3.75	2.40	3.97	5.00	4.00	4.60	4.33	3.85	3.35	4.01	3.17	3.41
BRAJA	3.82	4.17	5.50	3.80	4.09	4.00	3.77	3.79	4.00	2.40	4.14	4.83	5.20	3.80	4.33	3.70	3.59	4.06	3.09	3.00
RICPR	3.53	3.17	5.50	3.33	3.82	5.00	3.77	3.81	4.00	2.40	3.91	4.89	3.00	3.40	4.33	3.37	2.71	3.96	2.54	3.59
ORITS	2.76	1.50	3.50	2.67	4.09	5.00	3.23	3.42	3.50	3.20	3.17	4.56	3.40	2.80	4.33	2.30	2.47	4.07	2.91	2.79
EHRCH	1.59	1.50	4.75	1.95	4.45	2.50	3.14	3.67	4.00	3.40	3.77	4.28	3.20	4.00	2.33	2.70	2.65	3.88	2.60	2.03
EHRRI	1.91	2.00	3.75	1.62	3.73	3.75	4.32	3.40	2.75	2.20	3.26	4.17	3.20	3.20	2.00	3.04	3.06	3.97	2.77	2.50
HELPI	3.06	3.33	4.75	3.59	4.09	2.75	3.91	4.00	3.50	1.80	3.83	4.50	5.20	3.80	3.33	3.22	2.76	3.88	3.00	2.84
CHLTR	3.38	3.00	4.75	3.88	4.45	3.75	3.59	2.96	3.50	3.80	4.00	5.00	5.20	3.20	4.33	3.33	2.82	3.75	2.91	2.93
PORGI	2.97	3.67	4.75	3.70	2.64	2.50	3.45	3.90	4.00	2.80	3.83	5.00	4.00	3.80	3.33	3.56	2.88	3.57	3.20	2.62
TREPA	3.15	4.17	4.75	3.71	3.45	3.75	2.77	3.29	4.00	3.40	4.31	4.72	4.40	4.60	5.00	3.11	3.53	3.86	3.49	2.91
LEPIN	3.35	3.83	5.50	3.90	3.36	2.25	3.95	3.46	2.00	3.60	3.31	4.83	5.20	4.60	4.33	3.74	3.59	3.93	2.86	2.83
BORBU	3.32	3.33	4.75	3.47	4.09	3.25	3.86	3.02	4.25	2.60	3.54	4.78	3.80	3.20	3.00	3.11	3.06	3.71	2.86	3.10
THETH	3.09	4.67	4.75	4.02	3.18	3.75	4.27	3.60	4.50	4.40	4.29	4.22	3.40	3.80	3.67	3.41	2.71	3.83	3.49	3.71
BACSU	4.56	4.67	4.75	3.76	3.00	3.50	3.55	4.19	3.75	3.60	3.94	4.94	4.60	4.00	3.00	2.89	3.24	3.44	3.49	3.40
STAEP	3.94	4.17	1.75	3.66	2.91	3.75	3.18	3.85	3.50	3.60	2.83	4.89	4.00	3.40	3.67	3.30	2.94	3.31	3.54	3.17
LACLA	3.74	4.67	4.00	3.47	3.00	3.75	3.36	4.19	3.50	3.20	3.06	4.22	4.00	3.20	3.00	3.26	2.29	3.57	3.49	3.12
CLOPE	3.53	4.00	4.75	3.76	2.64	3.00	4.00	3.94	4.00	3.40	3.40	4.28	3.80	3.20	0.33	3.07	3.12	3.40	3.71	3.45
CLOTM	4.03	4.17	4.75	3.96	3.45	3.75	3.36	4.06	3.50	3.40	3.57	5.00	5.20	3.00	3.67	2.89	3.18	3.51	3.83	3.28
STRCO	3.00	3.83	5.50	3.67	2.45	4.00	3.64	3.06	4.00	2.60	3.49	4.61	2.80	3.80	5.00	3.48	2.59	3.67	2.91	3.48
MYCLE	3.38	4.67	4.25	3.74	2.27	3.00	3.55	2.96	4.00	2.60	3.77	4.67	4.00	3.80	3.00	3.30	1.82	3.78	3.06	3.17
MYCTU	2.94	2.83	4.50	3.28	2.91	3.50	3.00	2.29	4.00	3.60	3.60	3.83	4.60	2.60	2.00	2.96	1.12	3.39	2.69	2.38
MYCGE	3.06	4.00	2.50	2.84	2.73	4.00	2.82	2.42	3.50	1.00	2.80	2.83	0.60	n.a.	n.a.	2.37	1.88	2.83	1.94	2.69
SYNY3	3.00	4.17	4.75	3.36	3.18	3.25	3.68	3.56	3.50	3.20	3.74	4.28	4.20	4.60	3.33	2.78	1.82	3.75	3.40	2.79
SYNP7	3.32	4.67	5.50	3.39	2.82	4.00	4.05	3.54	3.50	4.40	3.51	4.94	5.20	3.40	1.00	3.11	3.41	3.50	2.60	3.10
PYRSA-Chl	2.41	3.83	3.50	3.05	2.82	4.00	3.73	3.04	3.75	2.20	3.40	4.61	5.20	4.60	2.00	3.15	3.06	3.18	2.80	3.21
GALSU-Chl	3.18	4.33	3.50	3.42	2.91	2.75	3.82	3.35	4.25	4.20	3.51	4.61	5.20	3.40	3.33	2.74	3.06	3.49	2.40	2.98
WHEAT-Chl	2.82	3.50	3.25	3.09	2.45	2.00	2.64	3.02	3.00	2.80	3.14	4.06	2.80	1.80	2.67	2.41	3.06	2.92	3.17	2.74
ARATH-Chl	2.38	4.17	4.00	3.33	2.91	3.75	3.77	3.40	n.a.	1.60	3.66	3.39	2.80	2.80	3.33	2.44	2.12	3.49	3.20	2.97
TRIVA-Hy	3.44	3.17	4.25	2.91	4.36	3.50	2.91	3.19	1.50	1.20	3.80	3.78	4.60	3.20	5.00	2.52	2.82	3.61	2.51	2.59
TRYCR-Mt	2.88	3.50	4.75	3.08	3.64	3.25	3.59	3.23	2.75	1.60	3.11	3.94	1.40	2.20	3.67	2.37	1.47	4.04	2.43	2.66
MAIZE-Mt	3.88	3.67	5.50	2.85	4.09	5.00	3.73	3.38	1.00	1.80	3.49	3.83	3.80	3.20	2.67	3.04	2.18	4.04	3.03	3.02
YEAST-Mt	3.21	3.17	4.25	3.16	4.45	2.50	3.86	3.21	2.75	1.60	3.69	3.50	4.00	2.40	2.67	2.30	2.47	3.78	3.06	2.48
HUMAN-Mt	3.71	2.50	3.50	3.22	4.45	3.75	3.50	3.17	2.00	2.20	2.91	3.72	3.00	2.20	3.00	1.78	2.18	3.67	3.20	2.81
HELVI-Mt	3.21	3.50	2.75	3.24	4.00	3.25	3.41	2.65	1.75	1.80	2.71	2.78	2.80	n.a.	0.67	2.07	3.00	3.43	2.71	2.72
PLAFG-Mt	2.68	4.17	4.00	2.08	2.64	2.25	1.77	2.19	4.00	1.80	2.51	n.a.	1.40	2.40	3.00	1.63	2.71	2.56	2.17	2.53

^a Average Conservation Index.^b Average Sequence Conservation.^c n.a. indicates a sequence not aligned.

conserved residue types are the aliphatics, which, however, are seldom individually conserved, followed by charged residues.

Highly conserved words (peptides)

Contiguous highly conserved residues (blue in Fig. 1) mostly relate to ATP/ADP or Mg^{2+} binding sites and to transition regions between structural Domains. Block 5 contains the highly conserved ATP/ADP binding sequence 30-T ϕ GPKG-35 (ϕ = L,I,V,M,A). Block 8 contains two highly conserved words corresponding to the ATP/ADP and Mg^{2+} binding segments 52-DGVTVAKEI-60 and 85-AGDGT²TATVL-95. Block 12 contains

the conserved word 191-EGMQFDRGYISPY-203, which traverses the transition region from Domain I to Domain A and includes several substrate binding sites mostly represented by aromatic residues. Two conserved words are contained in Domain A: 246-PLLI²AED-253 features a buried β -strand of five aliphatic residues; 275-AVKAPGFGDRRK-286 is enriched in charged residues, which in the *cis* conformation become exposed to the central cavity. Two highly conserved words occur in Block 22 and connect Domain A with Domain I, and Domain I with Domain E so that three of the four domain transition segments correspond to highly conserved words (see Fig. 1). Of the last two, 363-EKLQERLAKL

AGGVAVIKVG-382 includes several charged positions exposed to the central cavity in the *cis* conformation and a glycine doublet that separates the two domains; 402-ATRAVEEGIVPGGG-416 culminates with three consecutive perfectly conserved glycines that bind to the ATP/ADP cofactor. The protein conformation about the glycine triplet changes drastically between the *cis* and the *trans* conformation, suggesting that these residues may be crucial in coupling ATP hydrolysis to the rearrangement of domains distinguishing the two basic conformations.

Switch positions

Many switch positions surround the ATP/ADP binding site. For example, position 51, binding the third phosphate group of ATP, is exclusively occupied by N in Chl, Gram(+), Cyanobacteria, and PLAFG-Mt, and by K in all other sequences. Other switch positions correspond to indels between blocks of alignment. For example, between Blocks 10 and 11 all Rickettsiales, ZYMMO, and BORBU insert K or R, whereas all other sequences carry a non-specific hydrophobic residue.

Evolutionary conservation in relation to functional properties of the HSP60 proteins

We say that two residues or cofactors are contacting if the three-dimensional (3D) distance between their closest atoms is at most 4.0 Å, not counting residues part of the same chain separated by less than four amino acids.

ATP/ADP and Mg^{2+} binding sites are the most conserved (Table 3A)

The impressive evolutionary conservation (average CI = 0.84) of most of these residues attests to their functional importance. The predominant residues feature G (6 positions, 4 perfectly conserved), D (4 and 3), and T (4 and 2). Less conserved positions, like Y478, generally bind via backbone atoms. The nucleotide binding positions 51–53 at the start of Block 8 are deviant only in ZYMMO. All other positions are either strictly conserved or allow only conservative replacements.

Substrate binding positions are highly conserved (Table 3B–C)

Hydrophobic/aromatic positions Y199, Y203, F204, L234, L237, L259, V263, and V264 have been proposed as substrate binding since their mutation diminishes the substrate binding capacity of the GroEL complex (Fenton et al., 1994). These positions are highly conserved (average CI = 0.86) and in the alignment allow replacements only by other hydrophobic/aromatic residues. The overall picture depicts a conserved aromatic/hydrophobic environment in the central cavity of the *trans* conformation. A second set of residues have been proposed to bind to the substrate in a mini-chaperone with functional activity (Buckle et al., 1997). However, the mini-chaperone binding positions are much less conserved (average CI = 0.59) and in the multiple alignment are frequently represented by hydrophilic residues.

Table 3. Conservation of functional positions

A. ATP/ADP and Mg^{2+} binding residues											
T30 ^{a,b}	1.00	L31 ^{a,b}	0.75	G32 ^{a,b}	1.00	P33 ^{a,b}	1.00	K51 ^a	0.50	D52 ^b	0.94
D87 ^{a-d}	1.00	G88 ^{a,b}	1.00	T89 ^{a,b}	1.00	T90 ^{a,b}	0.92	T91 ^{a,b}	0.91	I150 ^a	0.91
D398 ^{c,e}	1.00	G414 ^{a,b}	1.00	G415 ^{a,b}	1.00	G416 ^{a,b}	1.00	I454 ^{a,b}	0.98	Y478 ^b	0.67
A480 ^{a,b}	0.79	A481 ^{a,b}	0.01	I493 ^{a,b}	0.90	D495 ^{a,b}	1.00			N479 ^{a,b}	0.55
B. Positions proposed to bind substrate by Fenton et al. (1994)											
Y199	0.88	S201	0.96	Y203	0.92	F204	0.69	L234 ^f	0.67	L237 ^f	1.00
V263	0.85	V264	0.70							L259	0.94
C. Positions proposed to bind substrate by Buckle et al. (1997)											
I230	0.53	E238	0.93	A241 ^f	0.32	E257 ^f	0.92	A260	0.44	T261 ^f	0.75
R268	0.84	T270 ^f	−0.02	L271	0.48					N265 ^f	0.96
D. Charged residues exposed to the central cavity in the <i>cis</i> ring											
K4	0.85	E41	0.50	K42	0.63	K58	0.76	E59	0.47	E61	0.69
K75	0.30	E76	0.50	K80	0.69	D83	0.82	E178	0.70	E179	0.65
D196	0.88	R197	0.83	D224	0.51	K225	0.63	K226	0.91	E252	0.77
E255	0.70	E257	0.92	K277	0.94	D283	0.86	R284	0.79	R285	0.81
E303	0.70	D304	0.65	K308	0.14	K327	0.96	D328	0.60	D359	0.46
E363	0.94	K364	0.91	E367	0.97	R368	1.00	K371	0.85	K380	0.47
K390	0.68	K393	0.96	E397	0.63	R404	0.61	E408	0.86	D523	0.59

^aResidues contacting ADP in asymmetric structure (1aon).

^bResidues contacting ATP in symmetric structure (1der).

^cResidues contacting Mg^{2+} in asymmetric structure.

^dResidues contacting Mg^{2+} in symmetric structure.

^ePosition 398 has minimum distance 4.3–4.6 Å from Mg^{2+} and was included here because discussed in Xu et al. (1997).

^fThese residues contact GroES in the asymmetric structure.

Conserved charged residues exposed to the central cavity (Table 3D)

Forty-eight charged residues are exposed to the central cavity in the *cis* conformation. In the *trans* conformation, 10 of these positions from Domain I and 13 from Domain A become buried within or between monomers, while the remaining 25 positions of Domain E remain exposed (Fig. 2A). Most of these positions are highly conserved in the alignment (Table 3D; Fig. 2B), supporting an important effect of charge in interactions with the substrate. Their average CI is 0.72 compared to an overall average CI 0.42 for other consensus charged residues. Only 17 of these charged positions engage in interactions between monomers in the *trans* conformation.

3D charge clusters

No linear (sequence) charge clusters were detected in HSP60 proteins. However, using the method of Karlin and Zhu (1996) for characterizing 3D (structural) clusters, we have identified charge clusters in the asymmetric *E. coli* structure (Table 4): (1) An *intra-monomer mixed charge cluster in the cis conformation* involves 16 charged residues, all but D291 and R345 exposed to the central cavity. (2) An *intra-monomer mixed charge cluster in the trans conformation* in Domain A embraces 13 residues, all but 4 exposed to the central cavity in the *cis* conformation. (3) An *inter-monomer mixed charge cluster in the trans conformation*

involves 20 residues that describe a charged interface between monomers predominantly engaging residues of opposite charge. In particular, favorable opposite-charge interactions feature the inter-monomer pairs R197-E386 (CI 0.83 and 0.94, respectively), R268-E257 (0.84, 0.92), and E386-R284 (0.94, 0.79). Most of these residues become exposed to the central cavity in the *cis* conformation.

Within ring monomer-monomer interactions

Fifty-five inter-monomer pair interactions are established in the *cis* and *trans* rings of the asymmetric structure and 66 in the symmetric structure, involving a total of 90 distinct positions. These are substantially conserved (average CI 0.60) but to a lesser extent than occurs for ATP/ADP and Mg^{2+} binding positions (0.87) or charged residues exposed to the central cavity (0.72). Paramount conserved interactions are listed in Table 5.

Contacts between Domains E of contiguous monomers are largely unchanged in the different conformations, consistent with the fact that major rearrangements in the structure involve only Domains A and I. Interactions between Domains E are characterized by a large number of hydrophobic contacts including on one face the conserved hydrophobic residues I6 (CI = 0.69), L73 (0.68), L513 (0.50), T517 (0.74), V521 (0.85), which relate on the opposite face to the conserved hydrophobic residues V39 (0.63), L40 (0.74), I49 (0.79), I60 (0.85). The interaction is also effected by conserved

Fig. 2. Charged residues exposed to the central cavity. **A:** Most of the charged residues from Domains I (green) and A (red), which are exposed to the central cavity in the *cis* conformation, become buried between monomers in the *trans* conformation. Charged residues from Domain E (blue) are exposed in both conformations. **B:** Conservation of charged residues exposed to the central cavity in the *cis* conformation.

Table 4. Structural charge clusters

Res ^a	CI ^b	Dom ^c	Intra		Inter ^d
			<i>cis</i>	<i>trans</i>	
D179	0.65	I			A
D185	0.16	I			A
R197	0.83	A			B
K207	0.55	A			A
E214	0.17	A			A
D224	0.51	A	•	•	
K225	0.63	A	•		
K226	0.91	A	•	•	
R231	0.32	A		•	
E232	n.a.	A		•	
E252	0.77	A	•	•	
D253	0.94	A	•		B
E255	0.70	A	•	•	
E257	0.92	A		•	
K277	0.94	A	•		B
D283	0.86	A	•		
R284	0.79	A	•		B
R285	0.81	A	•	•	
K286	0.81	A	•	•	
D291	0.95	A	•		
E303	0.70	A	•		
E304	0.65	A	•	•	
E308	0.14	A		•	
E310	0.62	A		•	
K311	0.19	A		•	
R322	0.40	A			A
R345	0.95	A	•		
D361	0.68	A			B
E363	0.94	A			B
K364	0.91	A			B
E367	0.97	A			B
R368	1.00	A	•		
K371	0.85	A			B
E386	0.94	I			A
E388	0.89	I			A
K390	0.68	I			A
K392	0.54	I			A
K393	0.96	I			A
E397	0.63	I			A

^aConsensus residue.^bConservation Index.^cStructural domain.^dA and B identify different monomers.

opposite charge interactions between K4 (0.81) and E61 (0.69), and between E518 (0.76) and R36 (0.70).

Inter-monomer contacts of Domains A and I are of special interest since they may be important for understanding the mechanism of the *cis* ↔ *trans* conformational switch. Between Domains A there are six inter-monomer contacts in the *cis* conformation, and 10 contacts in the *trans* conformation or in the symmetric structure. In the *cis* conformation, the conserved Y203 (0.92), V263 (0.85), and V264 (0.71) interact from one face with D304 (0.65), the unaligned position 305, and G306 (0.81). In the *trans* conformation, we observe only two highly conserved contacts at Domain A interfaces, involving E257 on one surface and R268 and G269 from the opposite surface. Thus, the weak interactions correspond-

Table 5. Within ring conserved inter-monomer contacts

Res	(CI)	Res	(CI)	Contact type (Str ^a)
Domain E		Domain E		
K4	(0.81)	E59	(0.47)	Charge (a,b,c)
K4	(0.81)	E61	(0.69)	Charge (a,b,c)
I6	(0.69)	I60	(0.85)	Hydrophobic (a)
M69	(0.30)	V39	(0.63)	Hydrophobic (a,c)
L73	(0.68)	V39	(0.63)	Hydrophobic (a,c)
L73	(0.68)	I49	(0.79)	Hydrophobic (b)
L513	(0.50)	I49	(0.79)	Hydrophobic (a,b,c)
T517	(0.74)	V39	(0.63)	Hydrophobic (a,b,c)
E518	(0.76)	R36	(0.70)	Charge (a,b,c)
V521	(0.85)	L40	(0.74)	Hydrophobic (a,b,c)
V521	(0.85)	I60	(0.85)	Hydrophobic (a,b)
K526	(0.42)	E63	(0.43)	Charge (c)
Domain A		Domain I		
R197	(0.83)	E386	(0.94)	Charge (b,c)
F281	(0.98)	M389	(0.42)	Hydrophobic (b,c)
D283	(0.86)	K181	(0.47)	Charge (b,c)
R284	(0.79)	E386	(0.94)	Charge (c)
R285	(0.81)	E386	(0.94)	Charge (c)
Domain A		Domain A		
E257	(0.92)	R268	(0.84)	Charge (c)
Domain E		Domain I		
V510	(0.86)	T385	(0.70)	Hydrophobic (a)
L513	(0.50)	V387	(0.46)	Hydrophobic (a)

^aContacts observed in the (a) *cis* ring or (b) *trans* ring of the asymmetric structure (Xu et al., 1997), or in the (c) symmetric structure (Boisvert et al., 1996). See Table 4 for other abbreviations.

ing to Domains A in the *cis* conformation change to stronger charge interactions in the *trans* conformation. Domain I contacts the contiguous Domain A in the *trans* and in the symmetric conformation, and the contiguous Domain E in the *cis* conformation. Within Domain I, positions 181–183 are in contact only in *trans*, whereas within the peptide 383–389 the same residues contact different regions of the neighboring monomer in the *cis* and in the *trans* conformation. Positions 181–183 are moderately (CI = 0.47 for K181) or weakly conserved (0.26 for S182, 0.23 for L183). However, position 183 is occupied either by a nonspecific hydrophobic residue or by an A or T residue, all capable of hydrophobic interactions with the conserved residue Y360 (CI = 0.79). K181 contacts D283 (0.86) from Domain A. From Domain I, particularly interesting is the behavior of E386, which in the *cis* conformation contacts E76 of Domain E (CI = 0.50) and in the *trans* conformation contacts R197 (0.83), R284 (0.79), and R285 (0.81) from Domain A.

Intra-monomer domain contacts (Table 6)

In the *cis* conformation Domain I functions as a lid for the ATP/ADP binding pocket of Domain E, with 28 pairwise interactions between domains, including several conserved hydrophobic contacts. Domain I also contacts Domain A via the sequence segments 169–174 and 366–370, respectively, but no strongly con-

Table 6. Conserved intra-monomer contacts between Apical, Intermediate, and Equatorial structural domains^a

Res	(CI)	Res	(CI)	Contact type (Str)
Domain A		Domain I		
M193	(0.77)	V276	(0.90)	Hydrophobic (b,c)
K322	(0.40)	E178	(0.70)	Charge (b,c)
K327	(0.96)	D83	(0.82)	Charge (b,c)
V333	(0.74)	T176	(0.73)	Hydrophobic (b,c)
V333	(0.74)	V378	(0.93)	Hydrophobic (b,c)
Domain A		Domain E		
E209	(0.48)	K58	(0.76)	Charge (b,c)
K327	(0.96)	D83	(0.82)	Charge (b,c)
Domain I		Domain E		
I150	(0.91)	I493	(0.90)	Hydrophobic (a)
I150	(0.91)	I494	(0.72)	Hydrophobic (a)
E391	(0.81)	K51	(0.50)	Charge (a)
E408	(0.86)	K498	(0.93)	Charge (a,b,c)
E409	(0.92)	K498	(0.93)	Charge (a,b,c)
E409	(0.92)	R501	(0.90)	Charge (a,b,c)

^aAbbreviations as in Table 5.

served interactions are observed. In the *trans* conformation, Domain I moves toward Domain A, resulting in more contacts between the two domains compared to the *cis* conformation (20 vs. 10) including some conserved hydrophobic interactions (V333 with V378 and T176) and charge interactions K322-E178, K327-D83, and K343-E191. Strikingly, in the *trans* conformation Domain A connects to Domain E via salt bridges highlighting the conserved pairs E209-K58 and K327-D83. Three oppositely charged interactions between Domains I and E are conserved in all conformations: E408-K498, E409-K498, and E409-R501. Mutations at the salt-bridge E409-R501 change the relative stability of the different GroEL conformations (Aharoni & Horovitz, 1997). To our knowledge, no mutational analyses have been done on E408 or K498.

Interactions between rings

Seven residues participate in interactions between rings in the symmetric structure (Table 7), including the conserved hydrophobic residue V464 (CI = 0.71), the moderately conserved charged residues K105 (0.63), E461 (0.51), and E467 (0.41), and the variable small residues A108, A109, S463. In the asymmetric structure, some rearrangement at the surface between rings produces new contacts, including a salt bridge between E461 (0.51) and R452 (0.34). The opposite charge character of this interaction is conserved in most eubacterial sequences but is lost in all organellar sequences precluding YEAST-Mt. Roseman et al. (1996) propose that the salt-bridge K105-E434 contributes to the allosteric switch between *cis* and *trans* conformations. However, E434 is poorly conserved and part of a long indel region, whereas the neighboring residue D435 is very conserved (0.71).

Positions contacting GroES (Tables 3, 8)

Primary contact positions from GroES feature the segment 22–31 including the conserved hydrophobic residues I25 (CI = 0.88), I26

Table 7. Inter-ring contacts^a

In asymmetric structure bound to ADP and GroES (Iaon, Xu et al., 1997)				In symmetric structure bound to ATP (Ider, Boisvert et al., 1996)			
<i>cis</i> ring		<i>trans</i> ring		<i>trans</i> ring		<i>trans</i> ring	
Res.	(CI)	Res.	(CI)	Res.	(CI)	Res.	(CI)
K105	(0.63)	A109 ^b	(0.59)	K105	(0.63)	A108 ^b	(0.35)
A111	(0.18)	X434	(n.a. ^c)	A108 ^b	(0.35)	K105	(0.63)
V438	(0.16)	X434	(n.a.)	K105	(0.63)	A109 ^b	(0.59)
E461	(0.51)	R452	(0.34)	A109 ^b	(0.59)	K105	(0.63)
E461	(0.51)	S463	(0.38)	E461	(0.51)	S463	(0.38)
S463	(0.38)	E461	(0.51)	S463	(0.38)	E461	(0.51)
S463	(0.38)	S463	(0.38)	S463	(0.38)	S463	(0.38)
S463	(0.38)	V464	(0.71)	S463	(0.38)	V464	(0.71)
V464	(0.71)	S463	(0.38)	V464	(0.71)	S463	(0.38)
V464	(0.71)	V464	(0.71)	V464	(0.71)	E467	(0.41)
V464	(0.71)	E467	(0.41)	E467	(0.41)	V464	(0.71)
E467	(0.41)	V464	(0.71)				

^aSee also Table 4 for abbreviations.^bBackbone contact.^cn.a. indicates positions not aligned.

(0.52), L27 (0.68), and A31 (0.72). These positions contact GroEL at the conserved hydrophobic residues L234 (0.67) and L237 (1.00), emphasizing the hydrophobic nature of the interactions between GroEL and GroES. However, mutational studies (Fenton et al., 1994) demonstrated that N265 is also vital to GroES binding, and

Table 8. Contact positions between GroEL and GroES^a

GroEL		GroES		Contact type
Res	(CI)	Res	(CI)	
L234	(0.67)	A22	(0.13)	Hydrophobic
L234	(0.67)	G23	(0.47)	
L234	(0.67)	I26	(0.52)	
L237	(1.00)	I26	(0.52)	
E238 ^b	(0.93)	G23	(0.47)	Hydrophobic
E238 ^b	(0.93)	I26	(0.52)	
A241	(0.32)	I25	(0.88)	
Q242 ^c	(0.29)	I25	(0.88)	
E257	(0.92)	A31	(0.72)	
T261	(0.75)	L27	(0.68)	
T261	(0.75)	P28 ^c	(0.54)	
T261	(0.75)	D29	(0.39)	
N265	(0.96)	L27	(0.68)	
X270	(0.02)	I25	(0.88)	
X270	(0.02)	L27	(0.68)	

^aSee Table 4 for abbreviations.^bPosition 238 modeled as Alanine in the *cis* conformation (Xu et al., 1997).^cBackbone contact.

indeed N265 has CI = 0.96. Most positions given in Buckle et al. (1997) are binding contacts to GroES in the asymmetric structure.

Discussion

Sequence positions essential to protein function and/or structure are conserved over a broad evolutionary range. Freely variable regions of a protein family are generally nonfunctional. Significant (nonrandom) sequence variations often reflect on functional differentiation in response to specific environments. Different species have different requirements and we expect that this also applies to HSP60 activities. Functional differentiation can often correspond to switch positions between evolutionary groups. Proteins of the extended HSP60 family (eubacterial GroEL, eukaryotic TCP1, and archaeal thermosome) are highly expressed (S. Karlin & J. Mrázek, unpubl. data), and this property may be related to the high mobility of the HSP60 genes within and between genomes that is observed in some evolutionary lineages. Duplication and horizontal transfer of HSP60 may promote functional adaptation and differentiation of these proteins. At the same time, these alterations could make HSP60 a poor indicator of phylogenetic relationships among their host organisms.

Conserved peptides and ATP hydrolysis

The ATP/ADP binding pocket constitutes the most evolutionarily conserved region of the protein, supporting the need for an exacting structural configuration to effect the ATPase enzymatic activity. All but two highly conserved words overlay ATP/ADP binding positions or correspond to transition regions between structural domains. Only the transition from Domain E to Domain I is weakly conserved, probably due to the fact that this sequence segment occupies an elongated loop with many side chains located at the outside surface. Prominent is the highly conserved sequence of the transition region from Domain I to Domain E, which continues into the totally conserved glycine triplet contacting ATP/ADP, which is central to a region of backbone flexibility between conformations. Sigler et al. (1998) propose that the process of release of GroES and substrate from the *cis* ring implies first ATP hydrolysis in the *cis* ring structure. The structural modification induced by ATP hydrolysis allows ATP to bind to the *trans* ring, which in turn causes GroES and substrate to be released. We hypothesize that the conserved domain-transition/ATP-binding segment is important in coupling ATP hydrolysis with GroES release.

Conserved glycine residues

In the multiple alignment of HSP60 proteins, the most conserved residue is glycine (G). It is well recognized that G residues can be a "filler" or "hinge" for backbone conformations that are generally excluded for other side chains. Conserved G multiplets are also observed in other ATPase protein families, like HSP70 (Karlin & Brocchieri, 1998) or RecA (Brendel et al., 1997). The Ramachandran plot of all perfectly conserved G residues (not shown) reveals that many of them belong to glycine-only backbone conformations and/or to hinge connections relevant to the *cis* ↔ *trans* structural reconfigurations. In particular, the perfectly conserved G192 and G375 are hinge positions at the transition regions between Domains I and A. The perfectly conserved G410, at the transition from Domain I to Domain E, corresponds to a glycine-only backbone structure. Another flexible position is G282, which is part of

a highly conserved loop of Domain A abundant in charged residues. Other conserved G residues, which undergo severe backbone torsion between *cis* and *trans* conformations, correspond to the ATP/ADP binding positions G32, G414, and G415.

Highly conserved charged residues

Charge allows a plethora of interactions, weak to strong, between molecules. Three-dimensional charge clusters commonly concentrate at the protein surface, in a crevice, or at the interface of multichain proteins (Zhu & Karlin, 1996). Mixed-charge clusters at inter-chain contacts can help stabilize quaternary structures and mediate protein-protein interactions. Electrostatic interaction is a mechanism for long-range and rapid effects that help orient the surfaces, thereby facilitating hydrophobic interactions. The abundance and conservation of charged residues lining the central cavity of the HSP60 complex in the *cis* conformation argue for a functional role of charge pertinent to substrate interactions. In the *trans* conformation, conserved charged residues belonging to Domains A and I are buried between monomers and are in good position to generate favorable electrostatic attractions. Inter-monomer interactions may depend on connections among charged residues (White et al., 1997). However, we observe many more conserved charged residues in the central cavity region of HSP60 in the *cis* conformation than conserved charge contacts in the *trans* conformation. Also, the charged residues exposed to the central cavity in Domain E, which do not become buried in the *trans* conformation, are as conserved as those from Domains I and A that become buried. These observations strengthen the view that charge has a special role in interacting with the substrate. Residues exposed to the central cavity interact with the substrate for about 15 s before this is released from the Anfinsen cage (Sigler et al., 1998). We propose that the highly conserved charged residues lining the central cavity of the GroEL complex favor the reconfiguration of the misfolded substrate into its native conformation. For example, they may interact ionically to help orient and position the substrate polypeptide in concert with hydrophobic binding residues or may play a part in "stretching" a misfolded substrate prior to release into the central cavity. Alternatively, charged residues can provide a strong hydrophilic environment that coerces formation of the hydrophobic core. In most of the HSP60 sequences, there is an excess of negative charge among these cavity positions. Given that proteins tend to be marginally acidic in most organisms (Stryer, 1995), the negative charge facing the central cavity may help to expel the substrate from the cavity after refolding.

Monomer interactions

The pronounced conservation at the interfaces between monomers and domains reflects their importance for effecting appropriate conformational movements in binding and releasing the substrate. Our observations suggest the following conclusions: (1) monomer-monomer interaction is mainly supported by contacts in Domain E with hydrophobic patches of conserved residues, hydrophilic interactions, and the conserved salt bridges of K4 with E61 and R36 with E518; (2) interactions of residues from Domains A or I are limited, but are more extensive in the *trans* conformation than in the *cis* conformation, suggesting that the *trans* conformation is inherently more stable. Charge interactions seem to destabilize the *cis* conformation (E386 with E76 in most eubacteria) and to stabilize the *trans* conformation (E386 with R197, R284 and R285,

and E257 with R268). Within each monomer, interactions between Domains I and A are also more extensive in the *trans* conformation and involve mainly conserved hydrophobic contacts and a few salt bridges; (3) few conserved inter-monomer charge interactions are observed in the *trans* conformation, suggesting that the conservation of charged residues exposed to the central cavity in the *cis* conformation may be more germane to their functional role in interacting with the substrate.

Interactions between rings

Interactions between rings in the *E. coli* structures do not include hydrophobic or charged pairs. Among all contacts the salt-bridges K105-D435 and E461-R452 are conserved in most eubacterial sequences but are absent in mammalian Mt sequences. It has been suggested that the HSP60 complex can only function as a double ring in *E. coli* but maintains chaperone activity as a single ring in mammalian Mt because of differences in the interaction mechanism of HSP60 with HSP10 (Nielsen & Cowan, 1998). An engineered version of the *E. coli* GroEL sequence with mutated inter-ring contact positions R452, E461, S463, and V464 produces, however, a single-ring complex which promotes refolding of substrates (Weissman et al., 1996). If the loss of the salt bridge capacity between positions 461 and 452 is responsible for the different structural properties of HSP60 in Mt sequences, we might predict that the eubacterial HSP60 sequences lacking the corresponding salt bridge could function as a single ring structure.

Functional differentiation among HSP60 proteins

Similarity assessments among HSP60 sequences are generally high. Some evolutionary groups diverge in particular regions of the protein. For example, the Rickettsiales *Orientia* and *Ehrlichia* diverge in Domain E. Domain I appears to be the most divergent in Gram(+) sequences. Domain A is the most variable among sequences, especially for *Orientia* and *Ehrlichia*'s, Mt sequences, and high G+C Gram(+). The proximity of many switch positions to the ATP/ADP binding pocket suggests that they may contribute to modulating the coupling mechanism of ATP binding/hydrolysis with binding and release of GroES and/or substrate. From this perspective, it could be revealing to experimentally test the significance of the switch observed between K and N at position 51. We observe enticing patterns of variability in the C-terminal portion of HSP60 sequences. The C-terminal tails in most of the sequences feature iterations of the triresidue GGM that are also found in many eukaryotic HSP70 C-terminal sequences. It has been determined that deletion of the C-terminal tail of HSP60 up to position V521 does not alter the viability of *E. coli* cultures (Burnett et al., 1994). The significance of these repetitive elements is unknown. Does the composition of these C-terminal elements reflect the functional differentiation of HSP60 sequences in different organisms or cellular compartments? For example, the sequence MYCTU from *Mycobacterium tuberculosis* as well as the sequence STRCO from *Streptomyces coelicolor* do not have any GGM elements in their C-terminal tail, but are enriched in histidines. Interestingly, a second paralog from *M. tuberculosis* possesses a perfect GGM element with several imperfect repetitions. This may suggest that the two Gram(+) sequences perform different functions in their respective organisms. A second GroEL gene is extant in *S. coelicolor* and we surmise that this sequence includes GGM C-terminal elements.

Mutational studies

Many mutational studies have been implemented on HSP60 proteins, including Fenton et al. (1994), Aharoni and Horovitz (1997), and White et al. (1997). For example, among conserved positions in monomer-monomer interactions, the mutant D283K maintains a wild-type phenotype; R197A associates with allosteric changes in the HSP60 complex (White et al., 1997); Y360E leads to reduced ATPase activity; F281D yields reduced ATPase activity and reduced substrate folding; V264S causes diminished ATPase activity and curtails GroES and polypeptide binding; Y203E causes no GroES binding, no polypeptide binding, but retains ATPase activity (Fenton et al., 1994); mutants of E386 cannot be purified (White et al., 1997). Other conserved positions involved in monomer-monomer interactions, R284, E257, R268, G269, T385, K364, G306, have not been examined. Conservation and mutation analyses offer independent tests on the structural or functional relevance of specific positions. The strong conservation observed among hydrophobic residues, which can alter the substrate binding capacities of HSP60 in mutational studies (Fenton et al., 1994), confirms their functional importance. On the other hand, among a different set of positions proposed to interact with the substrate from contacts observed in a synthetic mini-chaperone crystal structure (Buckle et al., 1997), the residues are much more variable, suggesting that such contacts have minor functional relevance. We propose all conserved positions and contacts listed in Tables 3–8 as attractive candidates for mutational studies.

Materials and methods

Multiple alignment

We used the ITERALIGN program (Brocchieri & Karlin, 1998); the method combines motif finding and dynamic programming analysis, based on the SSPA (Significant Segment Pair Alignment) protocol (Karlin et al., 1995; Brocchieri & Karlin, 1998). The multiple alignment consists of an array of core blocks, optimized using locally a dynamic alignment procedure, separated by unaligned residues. Aligned blocks can be independently studied as a potential functional/structural module. Each Block is summarized by a Consensus Sequence that best represents the composition of the alignment. A consensus residue is determined as the amino acid that maximizes the (weighted) average similarity score with respect to all residues in that column. In this way, each core block can be summarized by a vector of consensus residues. The quality of each aligned position within each Block is assessed by the Conservation Index (CI). Explicitly, the similarity matrix is normalized to $s^*(i,j) = s(i,j)/\sqrt{s(i,i)s(j,j)}$ such that for amino acid identity ($i = j$) the similarity score is 1.0. At each position of an alignment of N sequences, we calculate the Conservation Index (CI) as $[2/N(N-1)]\sum s^*(a,b)$, where the sum extends over all pairs of sequences for the given position in the alignment. The CI value at a perfectly conserved position is 1.0. At a position where predominantly similar amino acids appear (e.g., D with E) a positive CI, generally >0.4 , would be attained. A negative CI value reflects many nonconservative amino acid replacements. Averaging CI over the positions in a Block determines the Conservation Index of the Block. The Sequence Conservation within each Block is the average score of that sequence with the Consensus Sequence.

Structural clusters

All statistically significant three-dimensional residue clusters of amino acids with a particular property were investigated applying the method of Karlin and Zhu (1996).

Acknowledgments

We are happy to acknowledge helpful discussions with B.E. Blaisdell, A.L. Horwich, and P.A. Lund. S.K. is supported in part by NIH Grants 5R01GM10452-33 and 2R01HG00335-09.

References

- Aharoni A, Horovitz A. 1997. Detection of changes in pairwise interactions during allosteric transitions: Coupling between local and global conformational changes in GroEL. *Proc Natl Acad Sci USA* 94:1698–1702.
- Boisvert DC, Wang J, Otwinowski Z, Horwich AL, Sigler PB. 1996. The 2.4 Å crystal structure of the bacterial chaperonin GroEL complexed with ATP γ S. *Nat Struct Biol* 3:170–177.
- 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.
- Buckle AM, Zahn R, Fersht AR. 1997. A structural model for GroEL-polypeptide recognition. *Proc Natl Acad Sci USA* 94:3571–3575.
- Burnett BP, Horwich AL, Low KB. 1994. A carboxy-terminal deletion impairs the assembly of GroEL and confers a pleiotropic phenotype in *Escherichia coli* K-12. *J Bacteriol* 176:6980–6985.
- Fenton WA, Kashi Y, Furtak K, Horwich AL. 1994. Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* 371:614–619.
- Hartl FU. 1996. Molecular chaperones in cellular protein folding. *Nature* 381:571–580.
- Horovitz A. 1998. Structural aspects of GroEL function. *Curr Opin Struct Biol* 8:93–100.
- Karlin S, Brocchieri L. 1998. Heat shock protein 70 family: Multiple sequence comparisons, function, and evolution. *J Mol Evol* 47:565–577.
- Karlin S, Weinstock GM, Brendel V. 1995. Bacterial classification derived from RecA protein sequence comparison. *J Bacteriol* 177:6881–6893.
- Karlin S, Zhu Z-Y. 1996. Characterizations of distinctive residue clusters in protein 3D structures. *Proc Natl Acad Sci USA* 93:8344–8349.
- Netzer WJ, Hartl FU. 1998. Protein folding in the cytosol: Chaperonin-dependent and -independent mechanisms. *TIBS* 23:68–73.
- Nielsen KL, Cowan NJ. 1998. A single ring is sufficient for productive chaperonin-mediated folding in vivo. *Mol Cell* 2:93–99.
- Roseman AM, Chen S, White H, Braig K, Saibil HR. 1996. The chaperonin ATPase cycle: Mechanism of allosteric switching and movements of substrate-binding domains in GroEL. *Cell* 87:241–251.
- Sigler P, Zhaoxui X, Rye HS, Burston SG, Fenton WA, Horwich AL. 1998. Structure and function in GroEL-mediated protein folding. *Annu Rev Biochem* 67:581–608.
- Stryer L. 1995. *Biochemistry*. 4th ed. New York: WH Freeman and Co.
- Weissman JS, Rye HS, Fenton WA, Beechem JM, Horwich AL. 1996. Characterization of the active intermediate of a GroEL-GroES-mediated protein folding reaction. *Cell* 84:481–490.
- White HE, Chen S, Roseman AM, Yifrach O, Horovitz A, Saibil HR. 1997. Structural basis of allosteric changes in the GroEL mutant Arg197*Ala. *Nat Struct Biol* 4:690–694.
- Xu Z, Horwich AL, Sigler PB. 1997. The crystal structure of the asymmetric GroEL-GroES-(ADP)₇ chaperonin complex. *Nature* 388:741–750.
- Zhu Z-Y, Karlin S. 1996. Clusters of charged residues in protein three-dimensional structures. *Proc Natl Acad Sci USA* 93:8350–8355.