Phylogenetic Analysis of the Third Hsp70 Homolog in *Escherichia coli*; a Novel Member of the Hsc66 Subfamily and Its Possible Co-chaperone

Takeshi ITOH,¹* Hideo MATSUDA,² and Hirotada MORI¹*

Research and Education Center for Genetic Information, Nara Institute of Science and Technology, Ikoma 630-0101, Japan ¹ and Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, Toyonaka 560-8531, Japan ²

(Received 3 June 1999; revised 13 September 1999)

Abstract

Novel members of the highly conserved protein family, Hsp70, have been found in the complete sequences of several genomes. To elucidate a phylogenetic relationship among Hsp70 proteins of *Escherichia coli*, we searched all open reading frames derived from 13 complete genomes for Hsp70/actin-related proteins by the single-linkage clustering method. Phylogenetic analysis of this superfamily revealed that *E. coli* possesses at least three Hsp70 homologs (DnaK, Hsc66 and Hsc62). We found that Hsc62, which is the product of *hscC*, is a new member of the Hsc66 subfamily, and is specific to *E. coli*. The analysis also suggested that YegD of *E. coli* is closely related to the actin family, which consists of the actin, FtsA and MreB subfamilies. A further database search revealed that two *dnaJ* homologs, *ybeS* and *ybeV*, were located on the opposite strand near *hscC*. Consequently, *E. coli* seems to have three gene clusters composed of DnaK and DnaJ homologs.

Key words: complete genome; *Escherichia coli*; Hsp70; Hsc62; DnaK; DnaJ

Members of the 70-kD heat-shock protein (Hsp70) family are highly conserved molecular chaperones found in eu-bacteria, archaeabacteria and eukaryotes.¹ They function by preventing stress-induced damages, and mediate protein folding in an ATP-dependent process.² In *Escherichia coli*, two Hsp70 proteins, DnaK (*dnaK* gene product) and Hsc66 (*hscA* gene product) have been found,³⁻⁵ although expression of *hscA* is constitutive and not induced appreciably by heat shock.⁶ The YegD protein has also been reported to belong to this family, although only an amino-terminal fragment of the protein was determined at that time.⁷ Recently, the Japanese *E. coli* genome project determined *o170#1*⁸ and the complete sequence of *yegD* (*o356#2*).⁹ *o170#1*, located at 14.8 min on the genetic map, was suggested to encode a new member of the Hsp70 family as a result of a similarity search.⁸ The product of *o170#1*, named Hsc62, was indicated to function as a molecular chaperone.¹⁰ *o170#1* has been given the name *hscC* in SWISS-PROT.

The ATPase domain of the Hsp70 protein family shares common motifs with actin and sugar kinase protein families. An alignment based on a three-dimensional structural analysis of members of these three families showed that the domain includes five conserved regions, which displayed slight similarities at the amino acid level among the three families.¹¹ Of the *E. coli* proteins, FtsA, which is thought to be a member of the actin family,¹² and MreB¹³ belong to the Hsp70/actin/sugar kinase protein superfamily. The ATP-binding ability of phosphorylated FtsA in the cytoplasm has been confirmed experimentally,¹² based on a comparison of the conserved regions proposed by Bork et al.¹¹ Later, two phosphatases (Ppx and GppA) were also reported to be members of this superfamily.¹⁴ The ATPase domain of *Hsp70* is thought to have a crucial role in response to environmental changes in terms of not only evolutionary processes but also biological function, as shown in domain-exchanging study of yeast Hsp70 proteins.¹⁵ An evolutionary analysis of this conserved ATPase domain should facilitate the elucidation of the origin of the ubiquitous machinery for cell division and molecular chaperones.

We herein attempt to examine phylogenetic relationships among Hsp70 proteins including potentially new members. We classified proteins of the Hsp70/actin superfamily from 13 complete genomes on the basis of their amino acid sequence similarity, and then reconstructed their phylogenetic tree based on their ATPase domains.
Phylogenetic Analysis of Hsp70 Homologs [Vol. 6,

Table 1. Homologs of the Hsp70/actin superfamily in 13 complete genomes.

<table>
<thead>
<tr>
<th>Hsp70</th>
<th>actin/FtsA/MreB</th>
<th>unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aae</td>
<td>DnaK</td>
<td>FtsA, MreB</td>
</tr>
<tr>
<td>Bbu</td>
<td>DnaK, DnaK2</td>
<td>FtsA, MreB1</td>
</tr>
<tr>
<td>Bsu</td>
<td>DnaK</td>
<td>FtsA, MreB, MreBH, Mbl</td>
</tr>
<tr>
<td>Eco</td>
<td>DnaK, Hsc66, Hsc62</td>
<td>FtsA, MreB</td>
</tr>
<tr>
<td>Hin</td>
<td>DnaK, Hsc66</td>
<td>FtsA, MreB</td>
</tr>
<tr>
<td>Hpy</td>
<td>DnaK</td>
<td>FtsA, MreB</td>
</tr>
<tr>
<td>Mge</td>
<td>DnaK</td>
<td></td>
</tr>
<tr>
<td>Mpn</td>
<td>DnaK</td>
<td></td>
</tr>
<tr>
<td>Syn</td>
<td>DnaK1 (sll0170), DnaK2 (sll1932), DnaK3 (sll0058)</td>
<td>slr0086</td>
</tr>
<tr>
<td>Afu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mja</td>
<td>DnaK</td>
<td>MreB (MTH1024)</td>
</tr>
<tr>
<td>Mth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sce</td>
<td>SSA1, SSA2, SSA3, SSA4, SSB1, SSB2, GRP78, SSC1, YEL030w, YLR369w</td>
<td>ACT1, ARP1, ARP2, ARP3, LHS1, SSE1, SSE2, YHR064c</td>
</tr>
</tbody>
</table>

These analyses indicated that the E. coli Hsp70 family comprises at least three Hsp70 homologs, and that Hsc62 is a new member of the Hsc66 subfamily and is specific to E. coli. Furthermore, we found two new DnaJ homologs, one of which is adjacent to o170#1 in the opposite direction.

1. Clustering of the Hsp70/Actin/Sugar Kinase Superfamily

We searched all ORFs from 13 complete genomes for Hsp70/actin/sugar kinase proteins by using the single-linkage clustering method with the FASTA program. The complete genome sequences used in this study were those of Haemophilus influenzae (Hin), Mycoplasma genitalium (Mge), Synechocystis sp. (Syn), Methanococcus jannaschii (Mja), Mycoplasma pneumoniae (Mpn), Saccharomyces cerevisiae (Sce), Helicobacter pylori (Hpy), Escherichia coli (Eco), Bacillus subtilis (Bsu), Methanobacterium thermoautotrophicum (Mth), Archaeoglobus fulgidus (Afu), Borrelia burgdorferi (Bbu), and Aquifex aeolicus (Aae). These genome sequences and their annotations were obtained from the ftp site at ftp://ncbi.nlm.nih.gov/genbank/genomes/. Statistical significance of sequence similarity was evaluated by the z-value > 13 with 300 random shuffles of a query sequence.

About 60% of the E. coli K-12 W3110 genome has been determined by the Japanese E. coli genome-sequencing team, and the entire genome sequence of another strain (MG1655) of E. coli was determined by Blattner et al. independently of the Japanese project. However, the two substrains of E. coli have essentially the same genome sequence and identical results were obtained using genome data from either substrain.

Hypothetical members of the Hsp70/actin/sugar kinase superfamily, which were identified by the method described above, are shown in Table 1. Classification of the superfamily into three categories was accomplished according to the phylogenetic analysis (see below). As expected, proteins of the actin family (actin and FtsA) and the Hsp70 family were classified into the same superfamily, but not sugar kinases or phosphatases because of their low similarity at the amino acid sequence level. No Hsp70 protein was found in Methanococcus jannaschii or Archaeoglobus fulgidus, although this protein was thought to be ubiquitous in all life forms. Furthermore, M. jannaschii did not possess any member of the superfamily. Although it was expected that molecular chaperones played a crucial role in bacteria, these observations suggest that Hsp70 is not essential to these two thermophilic archaea (Methanococcus jannaschii and Archaeoglobus fulgidus, whose optimal growth temperatures are 85°C and 83°C, respectively), whereas another archaean Methanobacterium thermoautotrophicum, has Hsp70 and has an optimal growth temperature of 65°C. Further experimental analysis is required to elucidate the reason why the organisms lack the important protein.

In addition to DnaK, Hsc66, FtsA and MreB, which were previously classified as members of this superfamily, three new members, YegD, EutJ and Hsc62, were discovered among the E. coli proteins. Global alignments of the three functional domains of Hsp70 demonstrated that Hsc62 had a putative substrate-binding domain, and only a weakly-conserved carboxyl-terminal domain (Fig. 1). The carboxyl-terminal domain of DnaK is thought to play an important role in the interaction with DnaJ, while it was recently reported that amino-terminal interaction of DnaJ was localized to the ATPase domain of DnaK.
Figure 1. Amino acid identity between each of the domains of DnaK and the amino-acid sequences of different E. coli proteins as analyzed by the global-alignment program ALIGN.\textsuperscript{36} Black, hatched and shaded boxes indicate the ATPase, substrate-binding and carboxyl-terminal domains of DnaK, respectively. The amino acid length and identity of the corresponding region are shown in each box. Small black boxes under the ATPase domain of DnaK correspond to the five conserved regions defined by Bork et al.: P1, PHOSPHATE 1; Cl, CONNECT 1; P2, PHOSPHATE 2; A, ADENOSINE; C2, CONNECT 2.

Obviously, YegD, FtsA and MreB lacked the latter half of Hsp70, which contains the substrate-binding domain (Fig. 1). EutJ, which displays amino acid sequence similarity with the ethanolamine utilization protein EutJ in Salmonella typhimurium,\textsuperscript{34} was similar to MreB in both amino acid sequence and domain structure. In the following phylogenetic analysis, we discarded EutJ because it lacks the latter half of the ADENOSINE motif and the entire CONNECT 2 motif designated by Bork et al.\textsuperscript{11} The carboxyl-terminal region of Mrb, annotated by the Bacillus subtilis genome sequencing project,\textsuperscript{25} appears to include a frame-shift error. Thus, we used instead a sequence registered in SWISS-PROT (accession no. P39751).

2. Multiple-Alignment and Phylogenetic Analysis of the Hsp70/Actin Superfamily

Multiple sequence alignment was carried out by the CLUSTAL W program.\textsuperscript{40} Aligned sequences were improved manually with reference to the five motifs proposed by Bork et al.\textsuperscript{11} We reconstructed phylogenetic trees by using the neighbor-joining (NJ) method\textsuperscript{41} and the maximum likelihood (ML) method\textsuperscript{42,43} with amino acid sequences. CLUSTAL W\textsuperscript{40} and PROTMML included in the computer program package MOLPHY\textsuperscript{44} were used to assist the NJ and ML analyses, respectively. ML trees were obtained by the quick add search, using the Jones-Taylor-Thornton model\textsuperscript{45} of amino acid substitution, and 300 top ranking trees were retained (options -jf -q -n 300). Bootstrap values of the ML trees were calculated by analyzing 1000 replicates using the resampling of estimated log-likelihood (RELL) method.\textsuperscript{43}

Amino acid sequences of the proteins in Table 1 were aligned around the five motifs that constitute their ATPase domains. We also employed two more recently-determined Hsc66 sequences: Buchnera aphidicol\textsuperscript{a} (DDBJ/EMBL/Genbank accession no. AF008210) and Azotobacter vinelandii (AF010139). These five alignments were combined, and we used 132 sites which did not include any gaps. A phylogenetic tree was reconstructed by the ML method (Fig. 2). An almost identical tree was obtained by the NJ method (Fig. 2). An almost identical tree was obtained by the NJ method (data not shown), although some minor differences were found between the two trees due to the high divergence of the sequences employed. It was previously indicated that Hsp70 is remotely related to actin/FtsA.\textsuperscript{11} In fact, the superfamily could apparently be divided into two groups, Hsp70 and actin/FtsA/MreB, with respect to the phylogeny, while the phylogenetic positions of several proteins were ambiguous between the two groups.

According to the phylogenetic tree (Fig. 2) and the comparison of the gene structures (Fig. 1), YegD seems to be related to FtsA/MreB rather than to Hsp70. Alternatively, YegD may be a highly diverged member of the Hsp70 family which lost its substrate-binding
302 Phylogenetic Analysis of Hsp70 Homologs [Vol. 6, 

Figure 2. Phylogenetic tree based on ATPase domains reconstructed by the ML method. Numbers indicate bootstrap values of 600 or more for 1000 replicates. The scale for branch length is shown below the tree. Abbreviations for species are given in the text, except for *Azotobacter vinelandii* (Avi) and *Buchnera aphidicola* (Bap).

and carboxyl-terminal domains in the course of evolution. The cluster of the actin/FtsA/MreB family and YegD/slr0086 was significantly supported by a high bootstrap value. On the other hand, Hsc62 was apparently related to Hsp70 proteins (Fig. 2).

3. Novel Homologs of DnaK and DnaJ in *E. coli*

In general, if protein sequences have complex multidomain structures, single-linkage clustering analysis may result in the classification of mutually irrelevant sequences into one group. However, the domain structures of the proteins listed in Table 1 are rather simple, and we confirmed that all of them shared similarity with the ATPase domain of DnaK. We believe that ORF classification by the single-linkage clustering method is sufficiently effective, as has previously been shown,46,47 and can be used as an important first stage for further evolutionary analyses.

To find a protein closely related to Hsc62, we searched
Figure 3. (A) Multiple-alignment of amino-terminals of DnaJ homologs. All sequences except YbeS and YbeV were derived from SWISS-PROT: DNJ1.DROME, accession no. Q24133; CBPA.SALTY, O83916; DNAJ.AGRTU, P60018; DNAJ.FRATU, P48207; DNAJ.BACSU, P17631; DNJ2.RAT, P56402; DNJ2.HUMAN, P31689; HSJ1.HUMAN, P25686; DNJH.CUCSA, Q04960; DNJLARATH, P42825. Asterisks denote completely conserved sites, and colons denote highly conserved sites. (B) Genome organization around hscC. Homologous pairs are shown by the same patterns. The middle domain of YbeS or YbeV is slightly similar to YbeR or YbeU. Identity is given to each pair. An inverted repeat sequence followed by a T-stretch is illustrated as a putative terminator. BIME is a repetitive sequence composed of palindromic units.

Gupta and Golding found a gap in the N-terminal quadrant of archaeabacteria and Gram-positive bacteria, and suggested that this gap was a result of an insertion in other species. However, it seems difficult to account for this "insertion" event in a straightforward manner, because we found such "insertions" in FtsA and Hsc66 proteins, but not in Hsc62 and DnaK1-Bbu. This region may frequently accept insertions or deletions.

It will be interesting to search for proteins that interact with Hsc62. DnaJ, which is an Hsp40 family protein, does not stimulate the ATPase activity of Hsc62, whereas it does influence those of DnaK and Hsc66. In E. coli, dnaK and hscA are known to be co-transcribed with Hsp40 family members dnaJ and hscB, respectively. A database search for DnaJ homologs around hscC revealed that the products of ol69#12 (ybeS) and ol69#15 (ybeV) were slightly similar to well-conserved amino-terminal domains (J-domains) of DnaJ proteins (Fig. 3A). The products of ybeS and ybeV were highly homologous to each other (Fig. 3B). In addition, YbeR and YbeU were also slightly similar to the middle parts of YbeS and YbeV, but did not possess a J-domain. YbeT and YbeQ were also homologous to each other, and contained three and eight repeat units, respectively. Successive duplication events may have occurred in this region of the genome. Taking into consideration...

E. coli possesses three Hsp70 homologs, one Hsp70 (DnaK) and two Hsc66s (Hsc66 and Hsc62). However, no corresponding gene for Hsc62 has been found so far in other species, suggesting that Hsc62 is specific to E. coli. The optimal temperature of the ATPase activity of Hsc62 appears to be relatively close to that of Hsc66. This observation may reflect the evolutionary relationships and functional diversification among the three Hsp70 homologs of E. coli. DnaK1 of Borrelia burgdorferi could be an Hsc66 homolog but its phylogenetic location is unclear, as are several other sequences in the twilight zone between the Hsp70 and actin families.

a non-redundant protein database at the GenomeNet WWW site (http://www.genome.ad.jp/SIT/FASTA.html) for highly similar proteins to Hsc62. We then reconstructed an ML phylogenetic tree based on the conserved substrate-binding and carboxyl-terminal domains of these sequences. However, no protein sequenced so far was found to be monophyletic with Hsc62 (data not shown). This observation was in contrast to the ATPase domain phylogenetic tree (Fig. 2), where Hsc66 and Hsc62 were monophyletic. This was also the case for the NJ tree (data not shown). These results suggest that Hsc62 is a novel type of the Hsc66 protein, and that there exist at least two subfamilies in the Hsp70 superfamily: "classical" Hsp70 and Hsc66.

E. coli possesses three Hsp70 homologs, one Hsp70 (DnaK) and two Hsc66s (Hsc66 and Hsc62). However, no corresponding gene for Hsc62 has been found so far in other species, suggesting that Hsc62 is specific to E. coli. The optimal temperature of the ATPase activity of Hsc62 appears to be relatively close to that of Hsc66. This observation may reflect the evolutionary relationships and functional diversification among the three Hsp70 homologs of E. coli. DnaK1 of Borrelia burgdorferi could be an Hsc66 homolog but its phylogenetic location is unclear, as are several other sequences in the twilight zone between the Hsp70 and actin families.
the operon structures of the genes for the other DnaK and DnaJ proteins, it is quite possible that although hscC and the ancestor of ybeS and ybeV once cotranscribed, an inversion event later took place on hscC or the ybeS/ybeV ancestor. YbeS and YbeV are possible co-chaperones of Hsc62, although other DnaJ homologs, DjlA and CbpA, which are located at 1.2 and 23 min respectively and are not neighbored by any DnaK homolog, are also candidates for the co-chaperone of Hsc62. Consequently, *E. coli* was found to possibly possess three gene clusters of dnaK-dnaJ homologs: dnaK and dnaJ, hscA and hscB, and hscC and ybeS/ybeV.

There is a putative rho-independent transcription terminator for the upstream gene *ybeK* (Fig. 3B), which encodes a homolog of inosine-uridine nucleoside hydrolase. No promoter-like sequence could be found in the 5'-flanking region of hscC, while we can not exclude the possibility that hscC forms an operon with upstream genes. Further experimental studies of hscC is currently under way.

**Acknowledgments:** We thank T. Morimoto for helpful comments, and N. Takezaki and A. Stokes for critical reading of the manuscript. This work was supported in part by Grants-in-Aid (08283103 and 09276216) for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan.

**References**


