# Evolvability of Chaperonin Substrate Proteins

Emanuele Raineri[a,1], Paolo Ribeca[a,1], Luis Serrano[b], Tobias Maier[b]

[a] Department of Bioinformatics and Genomics, CRG, Dr Aiguader 88, ES-08003 Barcelona, Spain
[b] Department of Systems Biology, CRG, Dr Aiguader 88, ES-08003 Barcelona, Spain
[1] These authors contributed equally to this work

Corresponding Author:
tobias.maier@crg.es
Tel: +34933160186
Fax: +34933160099

March 23rd, 2009

## Abstract

Molecular chaperones ensure that their substrate proteins reach the functional native state, and prevent their aggregation. Recently, an additional function was proposed for molecular chaperones: they serve as buffers (capacitors) for evolution by permitting their substrate proteins to mutate and at the same time still allowing them to fold productively.

Using pairwise alignments of *E. coli* genes with genes from other gamma-proteobacteria, we showed that the described buffering effect cannot be observed among substrate proteins of GroEL, an essential chaperone in *E. coli*. Instead, we find that GroEL substrate proteins evolve less than other soluble *E. coli* proteins. We analyzed several specific structural and biophysical properties of proteins to assess their influence on protein evolution and to find out why specifically GroEL substrates do not show the expected higher divergence from their orthologs.

Our results culminate in four main findings: 1. We find little evidence that GroEL in *E.coli* acts as a capacitor for evolution *in vivo*. 2. GroEL substrates evolved less than other *E. coli* proteins. 3. Predominantly structural features appear to be a strong determinant of evolutionary rate. 4. Besides size, hydrophobicity is a criterion for exclusion for a protein as a chaperonin substrate.

# Introduction

Molecular chaperones are proteins which assist newly synthesized polypeptide chains to fold and mature to functional proteins. Additionally, under cellular stress conditions such as heat shock they are markedly over-expressed, to help prevent the aggregation of unfolded proteins. More recently, it has been proposed that chaperones carry out yet another function: they possess a buffer capacity against detrimental mutations, thereby functioning as a capacitor for evolution.

It was shown for Hsp90 both in *Drosophila* (1) and *Arabidopsis thaliana* (2) that impairing Hsp90 levels either genetically or pharmacologically leads to the appearance of an array of phenotypes. This is attributed to the fact that detrimental genetic polymorphisms, cryptic under conditions with regular chaperone levels, are phenotypically expressed once Hsp90 function is affected. The observed effect can be explained assuming that chaperones have a certain buffering capacitance (thus allowing substrate proteins to accumulate mutations and still reach the native state) whereas chaperone-independent proteins are more likely to misfold when acquiring mutations.

So far, the evidence supporting the hypothesis that chaperones function as a buffer for evolution mainly came from phenotypic observations (1, 2). Since evolution is based on genetic variance, and high phenotypic variability can be based on pleiotropic effects of a few mutations, we decided to investigate the buffering effect of chaperones at the genetic level. However, although satisfying genome data is published for Drosophila (3) with 12 fully sequenced species, no satisfying, unbiased list of Hsp90 substrate proteins is available to allow the investigation of the role played by chaperones for protein evolution in this model organism.

In fact, the concept of chaperones as buffers for evolution is not limited to Hsp90 alone. It was recently expanded to other classes of chaperones (4-6). In particular several biological and functional features of the bacterial GroEL/ GroES chaperonin system closely resemble those of the eukaryotic Hsp90:

1. GroEL, just like Hsp90 has a discrete set of substrate proteins, making them both specialist rather than generalist chaperones (7).
2. Both chaperones are essential for survival in their respective environment (8).
3. Hsp90 and GroEL/ES are abundant cellular proteins. Their levels can be decreased significantly by depletion or pharmacological impairment without affecting viability under permissive conditions (2, 9).
4. Both GroEL and Hsp90 bind metastable folding intermediates rather than nascent polypeptide chains (10, 11).
5. As observed for Hsp90 in Drosophila, overexpression of GroEL/ES in *E. coli* buffers against a fitness loss caused by deleterious mutations (6).

We therefore decided to base our study on GroEL and its well described substrate proteins in *Escherichia coli* (7), measuring evolutionary distances between chaperone substrate proteins and their orthologs in related bacteria and comparing them to evolutionary distances determined for proteins folding independently of GroEL (Table 1).

If GroEL worked as an evolutionary capacitor, we would expect GroEL substrates from *E. coli* and their orthologous partners in related organisms to show a greater sequence divergence than orthologous pairs of proteins folding inde-

pendently of GroEL. Instead we find lower sequence divergence between GroEL substrate proteins and their orthologs. Our data allows to attribute this finding predominantly to the particular structural composition of the GroEL substrate-set. In general, we do not dismiss the intriguing hypothesis that chaperones function as evolutionary capacitors. However, our data shows, that *in vivo* GroEL substrate proteins in *E. coli* evolved less than proteins folding independently of GroEL. Additionally, our analysis establishes hydrophobicity as a criterion for exclusion for a protein to be a chaperonin substrate.

# Results

## Evolvability

We based our analysis on a modified list of GroEL substrate proteins published by Kerner et al. (7). The authors isolated stabilized GroEL-GroES-substrate complexes and identified GroEL interacting proteins by mass spectrometry. A quantitative MS approach allowed the identified GroEL substrates to be sorted according to their abundance in complex with the chaperone, relative to their native levels in an *E. coli* cell lysate. The presented data is based on 204 GroEL substrate proteins (Table S2). The Selection criteria for this set of proteins are explained in the Methods section.

As a measure for evolutionary distance, we compared pairwise divergence between genes coding for GroEL substrate proteins of *Escherichia coli* and their orthologs in eight other gamma proteobacteria (*Buchnera aphidicola*, *Haemophilus influenzae*, *Pasteurella multocida*, *Photorhabdus luminescens*, *Salmonella typhimurium*, *Shigella flexneri*, *Vibrio cholerae*, *Yersinia pestis*) (Figure 1).

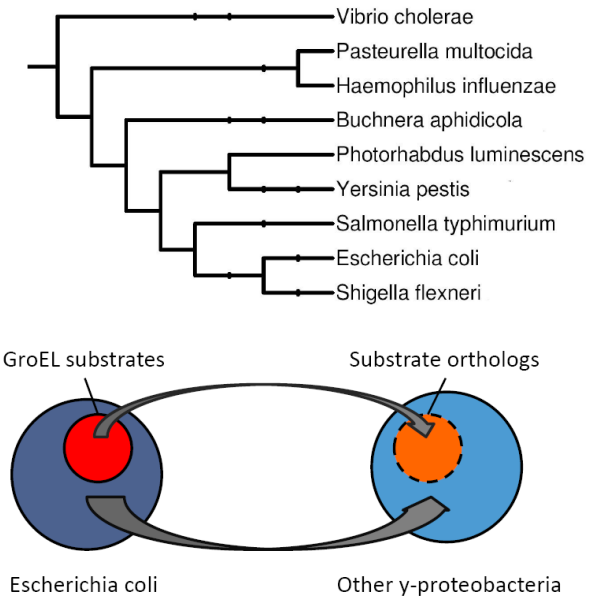



***Figure 1.*** *Evolutionary divergence between genes coding for GroEL substrate proteins and all mapped gene pairs in various gamma-proteobacteria.* ***A****: Organisms for which evolutionary distances to E. coli were calculated for this study.* ***B****: Cartoon depicting the approach for the calculation of evolutionary distances. Red/orange: GroEL substrate proteins and their orthologs. Blue/light blue: All orthologous protein pairs.*

To assess the evolutionary buffering capacity of GroEL, the calculated distances between pairs of orthologous genes coding for substrate proteins of the chaperone were compared to evolutionary distances for all mapped gene pairs for the respective organism (Figure 1). Evolutionary proximity of the analyzed organisms ensures that gene pairs selected on the basis of high sequence similarity correspond to orthologous proteins. Only gene pairs with at least 40% sequence identity were considered (Table 1).

Where available, the assigned gene pairs were verified by confirming that they share the same classification in terms of KEGG orthology classes (12).

| Organism | genome size | matched gene pairs | substrates compared | Evolutionary distance to E. coli | Ev. distance sub-strates to E. coli | Z-score | groEL similarity |
|---|---|---|---|---|---|---|---|
| Buchnera aphidicola | 574 | 495 | 46 | 0.091 | 0.081 | -1.7 | 94.50% |
| Haemophilus influenzae | 1657 | 1069 | 91 | 0.12 | 0.094 | -3.4 | 94.20% |
| Pasteurella multocida | 2015 | 1255 | 107 | 0.13 | 0.099 | -4.2 | 63.10% |
| Photorhabdus luminescens | 4683 | 1973 | 144 | 0.12 | 0.088 | -5.4 | 95.40% |
| Salmonella typhimurium | 4527 | 3187 | 184 | 0.074 | 0.05 | -4.2 | 99.60% |
| Shigella flexneri | 4445 | 3216 | 175 | 0.14 | 0.089 | -2.9 | 75.20% |
| Vibrio cholerae | 3835 | 1495 | 143 | 0.17 | 0.15 | -4.1 | 92.50% |
| Yersinia pestis | 4066 | 2310 | 159 | 0.12 | 0.093 | -4.4 | 96.40% |
| Escherichia coli | 4132 | 4132 | 204 | 0 | 0 | 0 | 100.00% |

***Table 1.*** *Analysed gamma-proteobacteria and organism specific parameters. Note that Z-scores for all compared organism pairs are negative, indicating a lower evolutionary distance between GroEL proteins and their orthologs in the respective organism, as compared to all mapped protein pairs.*

In accordance with the theory that chaperones serve as evolutionary buffers, chaperone dependence was expected to lead to a greater sequence divergence between orthologous genes coding for proteins stringently depending on GroEL for folding, as compared to orthologous gene pairs not coding for GroEL substrate proteins. Instead, we do not find evidence for higher evolutionary dynamics of GroEL substrates, but an opposite effect (Figure 2). In all eight analyzed pairs of organisms, genes coding for GroEL substrate proteins diverge less than their respective control sets of all mapped gene pairs for the tested organisms. (Table 1, Figure S1).

We reasoned that GroEL substrate proteins must possess specific properties, reversing the attributed buffering effect of the chaperone and hence accounting for the smaller genetic divergence. We analyzed parameters influencing the folding pathway (such as hydrophobicity) and structural properties of both GroEL substrates and proteins folding independently of GroEL. We also compared other possible determinants of evolutionary rate, namely expression level and essentiality (Table 2).

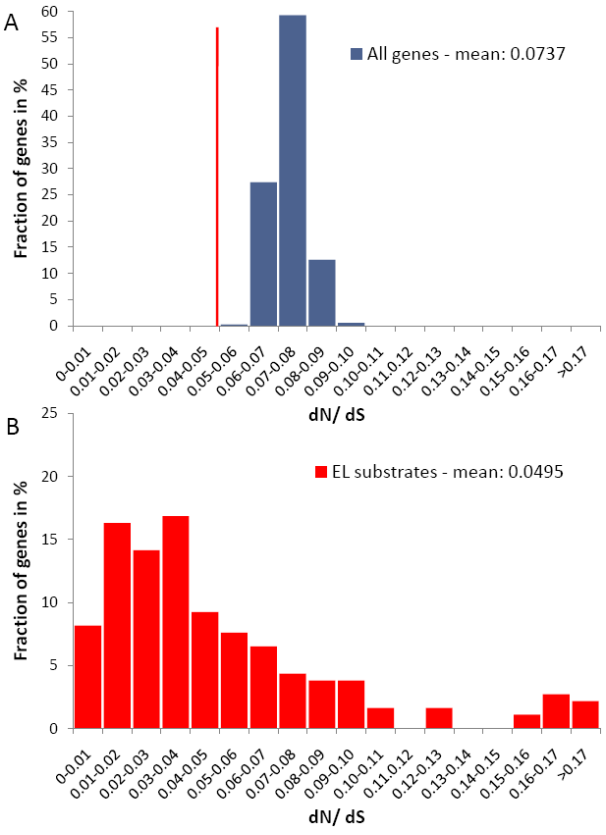



***Figure 2.*** *Genes coding for GroEL substrates evolve less than other E. coli genes.* ***A****: Distribution of evolutionary distances of 5000 random sub-sets of gene pairs between E. coli and S. typhimurium, each comprising 204 members. Red line: average evolutionary distance for Gro EL substrate proteins.* ***B****: Distribution of evolutionary distances of the 204 GroEL substrate genes between E. coli and S. typhimurium.*

We limited the analysis of these parameters to a comparison of *E. coli* with *S. typhimurium*. Several reasons suggested this pair as prime example of the analyzed gamma-proteobacteria:

1. Both organisms have comparable genome sizes (*E. coli*: 4132 genes, *S. typhimurium*: 4527 genes, Table 1)
2. Setting the sequence identity threshold to 90% still allowed us to confidently map 3316 gene pairs.
3. The respective groEL genes are 100% identical in their sequence (Table 1), suggesting that - although not a pre-requisite for this study - orthologous proteins to *E. coli* GroEL substrates also interact with the respective chaperonin of *S. typhimurium*.

The measured evolutionary distance (dN/dS) for the 204 mapped pairs of GroEL substrates between *E. coli* and *S. typhimurium* was 0.050 (Figure 2, Table 1). The average evolutionary distances of all mapped proteins pairs was 0.074 with a calculated Z-score of -4.2 (Figure 2, Table 1, Methods section).

## Essentiality

6.4 % of all genes in *E.coli* have so far been characterized as essential, according to the Pec Plus database (http://www.shigen.nig.ac.jp/ecoli/pecplus/index.jsp) (13). The average evolutionary distance of mapped gene pairs coding for essential proteins between *E. coli* and *S. typhimurium* is 0.056 with a Z-score of -3.88, compared to 0.074 for all mapped gene pairs. Among the genes coding for GroEL substrate proteins, 14.7%, or 30 out of 204 genes encode essential proteins. The 30 essential proteins among the GroEL substrate proteins (Table S1) have an average evolutionary distance of 0.040 (Z-score: -2.33), as compared to 0.050 for all chaperonin substrates.

The finding that essential genes are more conserved than non essential genes is in agreement with published data, mainly on yeast (14-16), but also on higher eukaryotes (17, 18), although contradicting data has also been published (19). Correcting for the enrichment of essential genes among GroEL substrates did not significantly alter the observed bias in evolutionary distance between GroEL substrates and control proteins (0.052 and 0.075 for GroEL substrates and all proteins, respectively).

## Expression level

Published data on yeast suggest a negative correlation between expression level and evolvability (19-21). We do not find a strong correlation between increased expression level and low evolvability in *E. coli* (Figure 3). The calculated Pearson coefficient is close to 0 (-0.055).

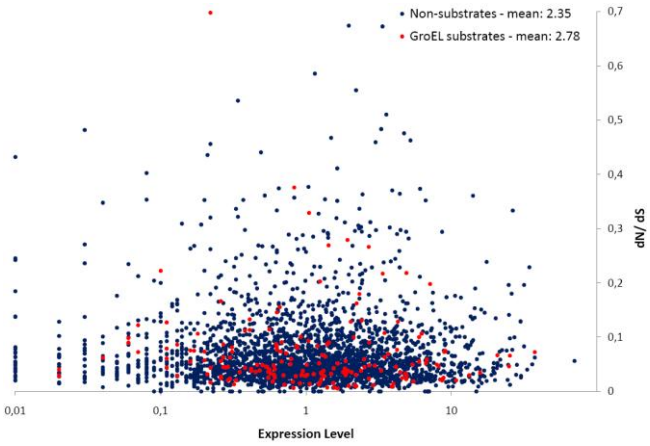



***Figure 3**. Expression level is not a major determinant for evolvability in E. coli. Genes coding for GroEL substrate proteins show slightly higher expression levels than genes coding for proteins not folding with GroEL. Red dots: GroEL substrate proteins. Blue dots: Proteins not interacting with GroEL for productive folding. For visibility reasons, the ordinate was shortened. Two data points for non-EL folders with evolvability values of 0.9365 and 1.303, respectively are missing in the graph.*

We used the GEO (Gene Expression Omnibus) database to analyze expression levels of *E. coli* genes (22). We found genes coding for GroEL substrate proteins to be expressed to a higher level (mean: 2.78, standard deviation 9.11, Z-score: 3.3), as compared to the genome wide average (mean: 2.35, standard deviation 4.53) (Figure 3).

The five GroEL substrate proteins with highest expression are PepQ, a proline peptidase; XylA, xylose isomerase; RimJ, ribsomal protein alanine acetyltransferase; GatY and GatZ, two D-tagatose-1,6-bisphosphate aldolase subunits (Table S1).

To account for a potential bias for low abundant proteins to be detected to a lesser extent by mass spectrometry, we repeated the analysis using published data-sets of experimentally detected *E. coli* proteins in proteomic studies (23-25). The mean abundance for this subgroup of 2019 proteins was 2.57 with a standard deviation of 5.14, practically leaving no significant expression level differences between *E. coli* substrate proteins and the control set.

## Protein Structure

Published data analyzing the correlation of protein structure and evolvability mainly focuses on the contact density, a measure for the designability of proteins. The contact density considers the fraction of buried amino acid residues in protein structures. It has been shown in some studies that evolvability correlates positively with the global contact density (26, 27). In contrast, other studies suggest that buried residues evolve less than amino acids exposed to the surface of a protein (28-30). We concluded that contact order is not a satisfying criterion to assess the relation of structure and evolution Protein structures can be categorized to higher detail by assigning them to hierarchical SCOP classes (31, 32). For this study, we based the analysis of structural properties of *E. coli* proteins and evolvability on SCOP class assignments.

A detailed structural analysis of GroEL substrate proteins revealed a bias among proteins stringently depending on GroEL towards certain SCOP fold classes (7). We tested the hypothesis that the SCOP class bias observed among GroEL substrates can account for the observation that GroEL substrate proteins evolve less than all other *E. coli* proteins, which fold in a chaperonin-independent manner.

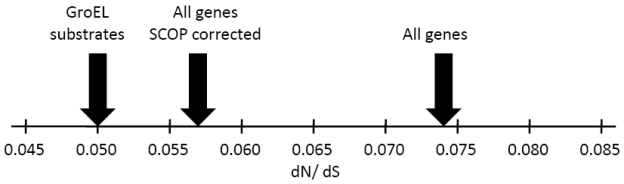



***Figure 4.*** *Predominantly structural reasons are responsible for the low evolvability of GroEL substrate proteins. Arrows indicate average evolutionary distances between E. coli and S. typhimurium for GroEL substrate proteins, all proteins, and all proteins with the same SCOP class distribution as for GroEL substrate proteins.*

We repeated the pair-wise alignment of genes between *E. coli* and *S. typhimurium*, reflecting the SCOP class distribution of the GroEL substrates in the control sets. While the evolvability of the random sub-sets shows a mean of 0.074, the SCOP class correction leads to a mean of 0.057 (Z-score: -2.93). This is considerably closer to the calculated mean considering the chaperonin substrate set alone (0.050, Z-score: -4.14, Figure 4). We therefore reason that specific structural properties of the GroEL substrate set are mainly responsible for the observed lower evolvability.

GroEL substrate proteins are highly enriched on the TIM barrel fold (SCOP class c.1) (7). We find that 33 of the 204 GroEL substrate proteins (16.2 %) adopt this fold, as compared to all other protein coding genes where only 2.9% (114 of all 3928 SCOP-annotated proteins) fold to TIM barrels. This structural class is an example of a fold with high sequence divergence, meaning that many different sequences acquire the same fold as native structure (33, 34). It has been speculated that highly designable structures (that is structures which are encoded by many different sequences) evolve rapidly (27), hence TIM barrel proteins in different organisms should show a high evolutionary divergence. What we find is an opposite effect: the average evolutionary distance between TIM barrels of *E. coli* and *S. typhimurium* is 0.061 (Z-score: -2.24), whereas the average distance between all proteins is 0.074. The enrichment of the GroEL substrates in TIM barrels and the low evolutionary divergence between the TIM barrels folding with GroEL (0.054) contributes significantly to the observation that GroEL substrates evolve less when compared to non-substrates.

## Hydrophobicity

We assessed the hydrophobicity of proteins using two algorithms: PEPWINDOW, which gives scores based on the Kyte-Doolitle index (35) and TANGO (36), measuring the aggregation propensity of proteins. For stringent GroEL substrates both algorithms independently showed a clear bias towards them being less hydrophobic than chaperonin-independent folders (proteins with predicted trans-membrane domains were excluded from the analysis). We further tested if the hydrophobicity scores of *E. coli* proteins are related to the evolvability of *E. coli* genes. The determined Pearson coefficient of 0.06 shows that there is no linear correlation between the calculated evolutionary distance for matched gene pairs and the respective hydrophobicity value of the corresponding *E. coli* proteins (Figure 5).

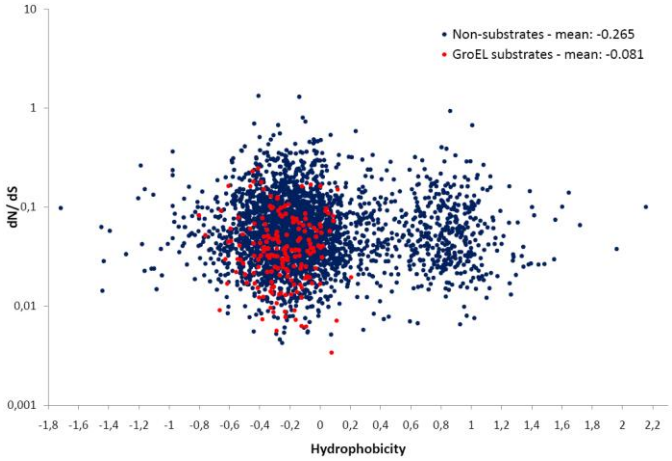



***Figure 5.*** *GroEL substrate proteins cluster at low hydrophobicity values and hydrophobicity does not correlate with evolvability in E. coli. Red dots: GroEL substrate proteins. Blue dots: Proteins not interacting with GroEL for productive folding.*

Our results suggest that GroEL substrate proteins on average are significantly less hydrophobic than other *E. coli* proteins, with hydrophobicity averages of -0.265 and -0.081, respectively and a Z-score of -3.4 (Figure 5). The supports of the distributions of the hydrophobicity values are also markedly different between GroEL substrate proteins (0.18) and proteins folding independent of GroEL (0.47). Figure 5 shows a large group of outliers in the control set with hydrophobicity values of above 0.3 (546 proteins). None of the GroEL substrate proteins reaches this value, with only three proteins in the substrate set having a Kyte-Doolitle score above 0.1 (Figure 5, Table S1). Whereas GroEL substrates cluster in a relatively small window between -0.7 and +0.1, other *E. coli* proteins have Kyte-Doolittle scores between -1.5 and +2.0. To account for a possible mass spectrometry induced bias, we repeated the analysis, using only experimentally identified *E. coli* proteins in

the control group (23-25). Using only experimentally determined proteins showed an even stronger statistical significant difference between the average values of the two data-sets (Z-score: -5.1).

## Discussion

We compared GroEL substrate proteins and chaperonin independent folders by evaluating biophysical, structural and physiological parameters with direct influence on evolution. Our analysis results in four findings (Table 2):

1. We find no evidence for GroEL functioning as capacitor for evolution in *E. coli in vivo*.
2. Instead, GroEL substrate proteins show a lower evolutionary distance to orthologous proteins, when compared to other proteins of *E. coli*.
3. Structural properties of GroEL substrate proteins account for their apparent lower evolvability.
4. Proteins folding with GroEL have closely clustered Kyte-Doolittle scores and are on average less hydrophobic than other cytoplasmic proteins of *E. coli*.

| Parameter | Evolvability | Substrate specific |
|---|---|---|
| Essentiality | ✓ | ● |
| Expression level | ✗ | ● |
| Structural properties | ✓ | ✓ |
| Hydrophobicity | ✗ | ✓ |

***Table 2.*** *Summary of the findings of the study. In E. coli essentiality correlates with low evolvability. There is a slight enrichment of essential proteins among GroEL substrates. Expression level does not correlate with evolvability. GroEL substrate proteins are slightly higher expressed than proteins not folding with GroEL. SCOP classes have different rates of evolution. GroEL substrates have a distinct distribution of SCOP classes. Hydrophobicity and evolvability are unrelated. GroEL substrates are less hydrophobic than proteins not folding with GroEL*

Our data suggest that, although GroEL was shown to buffer against deleterious mutations when overexpressed (6), appear to have functioned as a capacitor for evolution *in vivo*. In fact, we observe that GroEL substrate proteins in different gamma- proteobacteria evolve less than chaperonin-independent folders. This finding does not necessarily contradict the current view of how chaperone interaction during protein folding conveys a higher tolerance for mutations. We attributed the finding that GroEL substrate proteins evolve less mainly to the specific structural composition of the GroEL substrate set (Figure 4). To a lesser extent, the observed difference in evolvability is also due to an enrichment in essential proteins among GroEL-dependent folders. Protein abundance is not significantly different between GroEL substrate proteins and proteins folding independently of GroEL. This, together with the finding that GroEL can be depleted in *E. coli* without affecting viability (9), suggest that GroEL at native levels is not saturated with substrate proteins, even though they are expressed to a level above average. We therefore suggest that GroEL, at native levels, offers a fast, initial response mechanism to cellular stresses such as heat shock, before sigma-32 mediated over-expression of GroEL leads to additional available chaperone to accommodate for folding stress.

We found that GroEL substrate proteins are significantly less hydrophobic than other *E. coli* proteins. Hydrophobicity was shown not to correlate with the evolvability of proteins (Figure 5). The hydrophobicity scores of GroEL substrate proteins cluster in a much smaller range than that observed for chaperonin-independent folders. We believe that proteins with low hydrophobicity might not expose enough hydrophobic residues during their folding process to be recognized by the apical domains of the GroEL te-

tradecamer as a substrate protein. Proteins with a very high content of hydrophobic residues might undergo an initial collapse during their folding process, burying hydrophobic residues in the core of the protein, thereby removing them effectively from the pool of potential GroEL substrates. This could explain that virtually no GroEL substrate was identified with hydrophobicity values larger than 0.1 on the Kyte-Doolitle scale. These findings establish hydrophobicity as a criterion to exclude *E. coli* proteins from folding with the help of GroEL, similar to the observed size cut-off for GroEL substrates due to the limited capacity of the GroEL cavity (7, 37).

An ongoing debate is addressing the influence of translation fidelity on evolvability, related to an organism specific codon bias (20, 38). Since GroEL recognizes and folds its substrates post-translationally (39), an introduced codon bias due to synonymous mutations was not considered relevant for the analysis of the evolvability of chaperonin substrates.

Even though, based on our study, the hypothesis that chaperones function as evolutionary buffers does not seem to hold for the *in vivo* GroEL substrate proteins in *E. coli*, it remains an intriguing theory with much supporting data from different organisms (1, 2, 6, 40, 41). We suggest an experimental approach to validate the hypothesis in bacteria *in vivo*, employing modern high-throughput DNA sequencing techniques and quantitative proteomics: engineered *E. coli* mutator strains with regulatable levels of GroEL/ES expression could be grown for many generations expressing different levels of GroEL and GroES. A comparison of both the DNA sequences of the respective genomes and the sets of isolated and quantified GroEL substrate proteins before and after the growth experiments would allow one to draw conclusions on the effect chaperonin levels have on protein evolution. In addition, the analysis of acquired mutations of isolated and quantified GroEL substrates would potentially shed light on a yet unresolved question in biochemistry: what makes a protein a chaperone substrate. The publication of more complete chaperone substrate sets would allow additional bioinformatics-based evaluation of the buffering hypothesis for different organisms and different chaperone systems, including Hsp70 and Hsp90.

# Methods

## GroEL substrate-set

Kerner et al. distinguished three classes of GroEL substrate proteins and validated them experimentally (7). Class I proteins are abundant cellular proteins, which were also identified as GroEL interactors. Class II substrates are proteins which can use both GroEL and other chaperone systems for folding. Class III proteins stringently depend on GroEL for productive folding. For this study we excluded the identified substrate proteins belonging to class I, since we believe that these highly abundant cellular proteins do not represent typical GroEL substrates, but rather interact with GroEL on a stochastic basis. The presented data is hence based on 204 substrate proteins, comprising both class II and class III proteins. Taking into account only stringent GroEL substrates (84 proteins) does not significantly change the results of this study.

## Statistical setup

To assess the statistical significance of the differences between the GroEL substrates and the entire proteome we consistently adopted the following procedure, which hinges on the fact that the substrates are nothing but a particular subset of the proteome of cardinality n, and is

independent of the specific feature we are looking at.

1. We form N groups of cardinality n, randomly extracting them from the proteome. We always take N=5000.
2. For each of the subsets i, we compute its mean $m_i$, i=1 ... 5000. Notice that for large N we expect the $m_i$'s to be Gaussianly distributed, with the same average as the population's.
3. For each feature under examination, we compare the mean of the GroEL substrates $m_{GroEL}$ with the mean m and the standard deviation σ of the $m_i$'s. It is then possible to give a Z-score equal to $(m-m_{GroEL})/ \sigma$.

## Estimation of evolutionary distance

We downloaded the sequences of the proteins of all the organisms considered in this study from the KEGG database (ftp://ftp.genome.jp/pub/kegg/genes/organisms). For each transcribed gene of *E. coli* we then computed the corresponding closest gene in all other eight organisms. Closeness is defined by sequence identity calculated according to the Needleman-Wunsch algorithm (42) as implemented in NEEDLE (43). In this study we estimate the evolutionary distance as the ratio dN/dS.

To determine the evolutionary distance for each pair of genes assessed, we ran the software yn00 from the package PAML (44). This in turn implied first aligning the amino acid sequence of each gene product (for which we used MUSCLE (45)) then using this result to align the nucleotide sequences via TRANALIGN (43). This step was performed to make sure that the procedure introduces only gaps in multiples of three nucleotides, and hence does not produce artifact stop codons in the final result. Note the yn00 program returns a number of possible distances; we always use the one described in (46).

We noticed that in all cases where the number of substrates pairs stay in a suitable proportion with respect to the number of protein pairs found after imposing the similarity thresholds, the Z-scores computed as in Section "Statistical setup" remain significant (>3 σ) and negative (Table 1), pointing towards a diminished evolutionary rate of the GroEL substrates.

## Expression level

Expression data were taken from the database GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/ (22)) as our expression data source. Five data-sets with a sufficient number of expression profiles from wild type (WT) *E. coli* strains were randomly selected. The analyzed data-sets were: GDS680 (7 WT expression profiles, Z-score: 1.7), GDS1099 (all 15 expression profiles were considered here, since all were run on WT, although with different media/growing conditions, Z-score: 0.61), GDS2181 (6 WT expression profiles, Z-score: 3.3), GDS2768 (4 WT expression profiles, Z-score: 0.12) and GDS2825 (5 WT expression profiles, Z-score: 0.3). Expression values were averaged within each data-set, yielding five different expression estimates for each gene. The processed data-sets showed a very high correlation. The estimates obtained from data-set GDS2181 were used throughout this study.

The Z-scores computed as in Section "Statistical setup" point towards higher expression levels for GroEL substrates, although the significance of the result varies among the different datasets considered.

### Hydrophobicity

The hydrophobicity of the proteins in this study was assessed with PEPWINDOW, a software which gives scores based on the Kyte-Doolitle index (35) and TANGO (36), which essentially analyses the aggregation propensity of proteins. Both algorithms gave comparable results. Not to introduce a bias, in this study membrane proteins were excluded for the analysis of protein hydrophobicity. For those results, the Z-score computed as in Section "Statistical setup" is -5.1

### Protein structure

In the SCOP database version 1.73, 182 of the 204 GroEL substrate proteins had SCOP classes assigned (and 2644 of the 3928 annotated *E.coli* proteins in total). They fall into the following classes: a: 10, b: 13, c: 109, d: 41, e: 6, f: 2, g: 1.

To assess the contribution of structural properties to the evolutionary rates of GroEL substrates, we extracted randomly N=1000 samples from the *E.coli* proteome. To build each sample, we choose randomly 10 proteins of class a, 13 of class b, etc... as to reflect the structural properties of the GroEL substrates. We than applied the statistical procedure described above to the sample distribution. We further extracted only the TIM barrel proteins, to compare them with the rest of the proteome as described in the main text.



## Acknowledgements

TM is funded by an EMBO long-term fellowship. We would like to thank Ben Lehner for critically reading and correcting the manuscript.

## Footnotes

ER and PR designed and performed the computational analysis and analyzed the data. LS contributed intellectually and corrected the manuscript. TM had the idea, planned the study, prepared the figures and wrote the manuscript.

The authors declare no conflict of interest in connection with the manuscript.

# References


1. Rutherford SL & Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. *Nature* 396:336-42.

2. Queitsch C, Sangster TA & Lindquist S (2002) Hsp90 as a capacitor of phenotypic variation. *Nature* 417:618-24.

3. Clark AG, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203-18.

4. Tomala K & Korona R (2008) Molecular chaperones and selection against mutations. *Biol Direct* 3:5.

5. Sangster TA, Lindquist S & Queitsch C (2004) Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. *Bioessays* 26:348-62.

6. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF & Barrio E (2002) Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature* 417:398.

7. Kerner MJ, et al. (2005) Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell* 122:209-20.

8. Young JC, Moarefi I & Hartl FU (2001) Hsp90: a specialized but essential protein-folding tool. *J Cell Biol* 154:267-73.

9. Kanemori M, Mori H & Yura T (1994) Effects of reduced levels of GroE chaperones on protein metabolism: enhanced synthesis of heat shock proteins during steady-state growth of Escherichia coli. *J Bacteriol* 176:4235-42.

10. Nathan DF, Vos MH & Lindquist S (1997) In vivo functions of the Saccharomyces cerevisiae Hsp90 chaperone. *Proc Natl Acad Sci U S A* 94:12949-56.

11. Hayer-Hartl MK, Ewbank JJ, Creighton TE & Hartl FU (1994) Conformational specificity of the chaperonin GroEL for the compact folding intermediates of alpha-lactalbumin. *Embo J* 13:3192-202.

12. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.

13. Kato J & Hashimoto M (2007) Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol* 3:132.

14. Zhang J & He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22:1147-55.

15. Wall DP, et al. (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A* 102:5483-8.

16. Hirsh AE & Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* 411:1046-9.

17. Liao BY, Scott NM & Zhang J (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol* 23:2072-80.

18. Larracuente AM, et al. (2008) Evolution of protein-coding genes in Drosophila. *Trends Genet* 24:114-23.

19. Pal C, Papp B & Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927-31.

20. Drummond DA, Raval A & Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327-37.

21. Drummond DA, Bloom JD, Adami C, Wilke CO & Arnold FH (2005) Why highly ex-

pressed proteins evolve slowly. *Proc Natl Acad Sci U S A* 102:14338-43.

22. Barrett T, et al. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35:D760-5.

23. Gevaert K, et al. (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins. *Mol Cell Proteomics* 1:896-903.

24. Corbin RW, et al. (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci U S A* 100:9232-7.

25. Ishihama Y, et al. (2008) Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* 9:102.

26. Zhou T, Drummond DA & Wilke CO (2008) Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Evol* 66:395-404.

27. Bloom JD, Drummond DA, Arnold FH & Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol* 23:1751-61.

28. Bustamante CD, Townsend JP & Hartl DL (2000) Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Mol Biol Evol* 17:301-8.

29. Dean AM, Neuhauser C, Grenier E & Golding GB (2002) The pattern of amino acid replacements in alpha/beta-barrels. *Mol Biol Evol* 19:1846-64.

30. Goldman N, Thorne JL & Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445-58.

31. Murzin AG, Brenner SE, Hubbard T & Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-40.

32. Andreeva A, et al. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36:D419-25.

33. Wierenga RK (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett* 492:193-8.

34. Copley RR & Bork P (2000) Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol* 303:627-41.

35. Kyte J & Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105-32.

36. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J & Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302-6.

37. Sigler PB, et al. (1998) Structure and function in GroEL-mediated protein folding. *Annu Rev Biochem* 67:581-608.

38. Drummond DA & Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341-52.

39. Hartl FU & Hayer-Hartl M (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 295:1852-8.

40. Rutherford S, Hirate Y & Swalla BJ (2007) The Hsp90 capacitor, developmental remodeling, and evolution: the robustness of gene networks and the curious evolvability of metamorphosis. *Crit Rev Biochem Mol Biol* 42:355-72.

41. Yeyati PL, Bancewicz RM, Maule J & van Heyningen V (2007) Hsp90 selectively modulates phenotype in vertebrate development. *PLoS Genet* 3:e43.

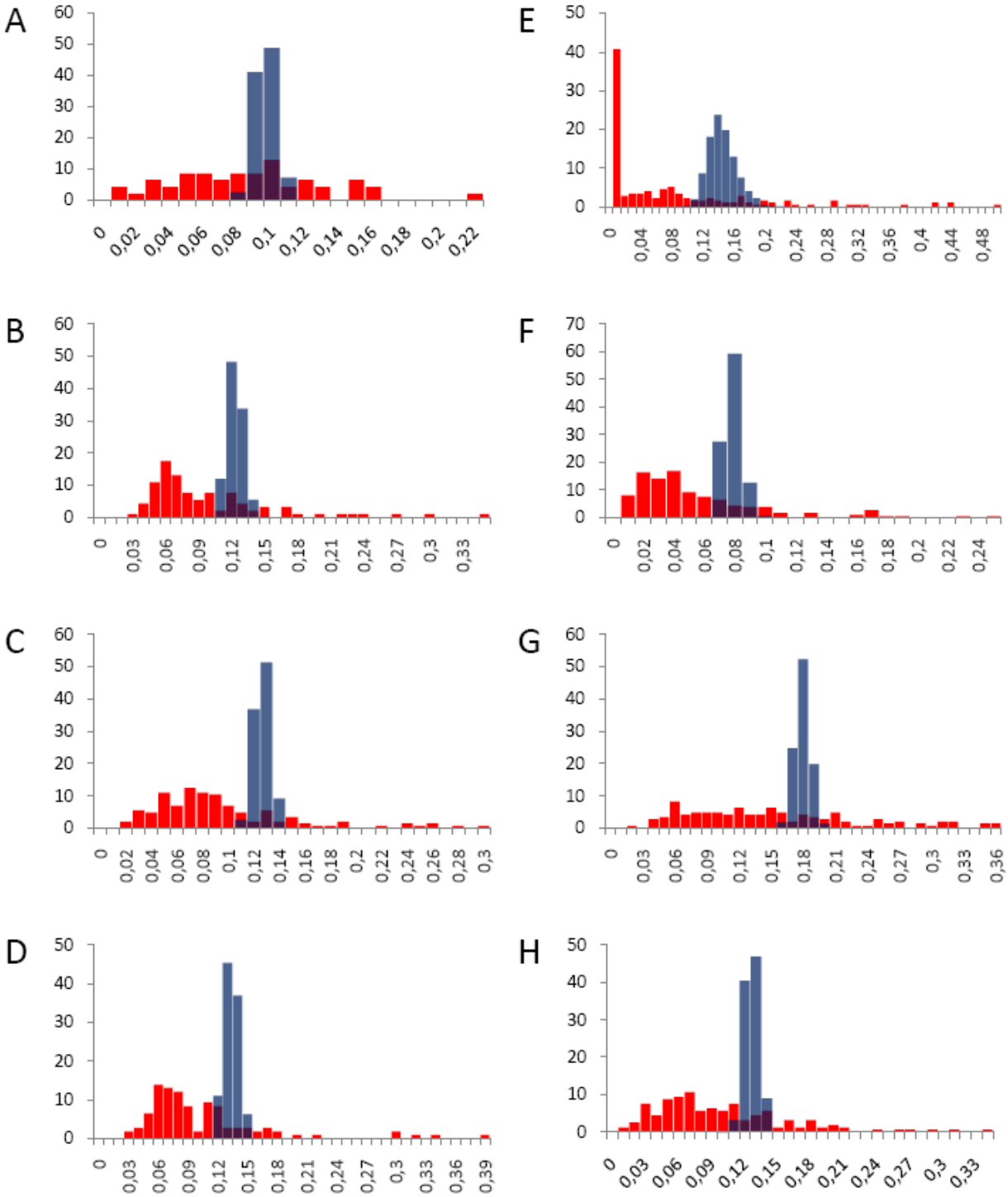
42. Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443-53.

43. Rice P, Longden I & Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276-7.

44. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-91.

45. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

46. Yang Z & Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32-43.




***Supplementary Figure 1***. *Histograms of evolutionay distances between GroEL substrates and all proteins for pairs of E. coli and different gamma-proteobacteria. Red bars: Distribution of evolutionary distances of GroEL substrate genes. Blue bars: Distribution of evolutionary distances of 5000 random sub-sets of gene pairs, each comprising 204 members. The abscissa shows dN/dS, the ordinate the fraction of genes of the data sets in % falling in each bin.* ***A****: Buchnera aphidicola;* ***B****: Haemophilus influenzea;* ***C****: Photorhabdus luminescens;* ***D****: Pasteurella multocida;* ***E****: Shigella flexneri;* ***F****: Salmonella typhimurium;* ***G****: Vibrio cholerea;* ***H****: Yersinia pestis. Averages for respective evolutionary distances are given in Table 1. Note that for illustrative purposes Figure 2 A/B and Figure S1 F are identical.*

| **GroEL substrate proteins with high evolutionary distance** | | | |
|---|---|---|---|
| Gene name | Protein name | Uniprot ID | Function |
| b0178 | SKP | P0AEU7 | Chaperone protein skp |
| b2095 | GatZ | P0C8J8 | D-tagatose-1,6-bisphosphate aldolase subunit |
| b0776 | BioF | P12998 | 8-amino-7-oxononanoate synthase |
| b3092 | Uxac | P0A8G3 | Uronate isomerase |
| b1106 | ThiK | P75948 | Thiamine kinase |
| **Essential GroEL substrate proteins** | | | |
| Gene name | Protein name | Uniprot ID | Function |
| b0154 | HemL | P23893 | Glutamate-1-semialdehyde 2,1-aminomutase |
| b0181 | LpxA | P0A722 | Acyl-UDP-N-acetylglucosamine O-acyltransferase |
| b0185 | AccA | P0ABD5 | Acetyl-coA carboxylase carboxyl transferase alpha |
| b0369 | HemB | P0ACB2 | Delta-aminolevulinic acid dehydratase |
| b1093 | FabG | P0AEK2 | 3-oxoacyl-[acyl-carrier-protein] reductase |
| b1204 | Pth | P0A7D1 | Peptidyl-tRNA hydrolase |
| b1207 | KprS | P0A717 | Ribose-phosphate pyrophosphokinase |
| b1215 | KdsA | P0A715 | 2-dehydro-3-deoxyphosphooctonate aldolase |
| b1719 | ThrS | P0A8M3 | Threonyl-tRNA synthetase |
| b2153 | FolE | P0A6T5 | GTP cyclohydrolase 1 |
| b2231 | GyrA | P0AES4 | DNA gyrase subunit A |
| b2478 | DapA | P0A6L2 | Dihydrodipicolinate synthase |
| b2533 | SuhB | P0ADG4 | Inositol-1-monophosphatase |
| b2607 | TrmD | P0A873 | tRNA (guanine-N(1)-)-methyltransferase |
| b2608 | RimM | P0A7X6 | Ribosome maturation factor rimM |
| b2925 | FbaA | P0AB71 | Fructose-bisphosphate aldolase class 2 |
| b2942 | MetK | P0A817 | S-adenosylmethionine synthetase |
| b3019 | ParC | P0AFI2 | DNA topoisomerase 4 subunit A |
| b3168 | InfB | P0A705 | Translation initiation factor IF-2 |
| b3251 | MreB | P0A9X4 | Rod shape-determining protein mreB |
| b3256 | AccC | P24182 | Biotin carboxylase |
| b3433 | Asd | P0A9Q9 | Aspartate-semialdehyde dehydrogenase |
| b3463 | FtsE | P0A9R7 | Cell division ATP-binding protein ftsE |
| b3650 | SpoT | P0AG24 | Guanosine-3',5'-bis (di-P) 3'-pyrophosphohydrolase |
| b3783 | Rho | P0AG30 | Transcription termination factor rho |
| b3850 | HemG | P0ACB4 | Protoporphyrinogen oxidase |
| b3865 | EngB | P0A6P7 | Probable GTP-binding protein engB |
| b3982 | NusG | P0AFG0 | Transcription antitermination protein nusG |
| b3987 | RpoB | P0A8V2 | DNA-directed RNA polymerase subunit beta |
| b3988 | RpoC | P0A8T7 | DNA-directed RNA polymerase subunit beta' |
| **Highly expressed GroEL substrate proteins** | | | |
| Gene name | Protein name | Uniprot ID | Function |
| b3847 | PepQ | P21165 | Proline peptidase |
| b3565 | XylA | P00944 | Xylose isomerase |
| b1066 | RimJ | P0A948 | ribsomal protein alanine acetyltransferase |
| b2096 | GatY | P0C8J6 | D-tagatose-1,6-bisphosphate aldolase subunit |
| b2095 | GatZ | P0C8J8 | D-tagatose-1,6-bisphosphate aldolase subunit |
| **Most hydrophobic GroEL substrate proteins** | | | |
| Gene name | Protein name | Uniprot ID | Function |
| b0154 | HemL | P23893 | Glutamate-1-semialdehyde 2,1-aminomutase |
| b2091 | GatD | P0A9S3 | Galactitol-1-phosphate 5-dehydrogenase |
| b2107 | KprS | P0A717 | Ribose-phosphate pyrophosphokinase |

***Supplementary Table 1****. Lists of GroEL substrate proteins and attributed properties: 1. High evolutionary distance 2. Essentiality; 3. High expression level; 4. High hydrophobicity.*