The cumulative impact of chaperone mediated protein-folding during evolution

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

David Bogumil

aus Düsseldorf

Kiel, im August 2013

aus dem Institut für Allgemeine Mikrobiologie der Christian-Albrechts-Universität zu Kiel

Die hier vorgelegte Dissertation habe ich eigenständig, ohne unerlaubte Hilfe und unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Teile der kumulativen Dissertation sind bereits veröffentlicht bzw. zur Veröffentlichung eingereicht. Diese sind als Solche gekennzeichnet. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Kiel, den

David Bogumil

Referentin: Prof Dr. Tal Dagan

Koreferent: Prof. Dr. William F. Martin

Tag der mündlichen Prüfung: 27.09.2013

Zum Druck genehmigt: Kiel , den 27.09.2013

Der Dekan

Content

1 Abstract 1

2 Zusammenfassung 2

3 Introduction 3

3.1 Chaperone-mediated protein folding in the three domains of life 3

3.2 Chaperone-mediated protein unfolding 9

3.3 Chaperone-mediated protein translocation 9

3.4 Protein misfolding and fitness 10

3.5 Chaperone-mediated protein folding and fitness 14

3.6 Chaperone-mediated folding and protein evolution 17

3.7 The evolution of protein interaction with chaperones 21

3.8 Thematic content 24

4 References 26

5 Publications 42

5.1 Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes 42

5.2 Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate 50

5.3 Imprints of early endosymbiosis within the chaperone interactions network in yeast 59

6 Acknowledgements 82

1 Abstract

Molecular chaperones support protein folding and unfolding along with assembly and translocation of protein complexes. Chaperones have been recognized as important mediators between organismal genotype and phenotype as well as important maintainers of cellular fitness under environmental conditions that induce high mutational loads. This thesis presents recent studies revealing that the folding assistance supplied by chaperones is evident in genomic sequences, thus implicating chaperone-mediated folding as an influential factor during protein evolution. Furthermore the evolution and the symbiogenic origin of the eukaryotic chaperone repertoire are elucidated. Protein interaction with chaperones ensures a proper folding and function, yet an adaptation to obligatory dependence on such assistance may be irreversible, representing an evolutionary trap. Correlation between chaperone requirement and protein expression level indicate that the evolution of substrate-chaperone interaction is bounded by the required substrate abundance within the cell. Accumulating evidence suggests that the utility of chaperones is governed by a delicate balance between their help in mitigating the risks of protein misfolding and aggregate formation on the one hand, and the slower rate of protein maturation and the energetic cost of chaperone synthesis on the other.

2 Zusammenfassung

Chaperone unterstützen sowohl die Proteinfaltung und Entfaltung als auch die Translokation von Proteinkomplexen. Sie spielen eine wichtige Rolle als Mittler zwischen Genotyp und Phänotyp, indem sie zur Aufrechterhaltung der *„fitness“* nach dem Darwin’schen Evolutionsmodell bei hoher Mutationsbelastung beitragen. Diese Dissertation zeigt anhand kürzlich veröffentlichter Studien, dass die Auswirkungen der Chaperon-unterstützten Proteinfaltung in genomischen Sequenzen messbar sind und die Chaperon-unterstütze Proteinfaltung daher einen Einfluss auf die Proteinevolution hat. Des Weiteren wird die evolutionäre Entwicklung des eukaryotischen Chaperon-Repertoires im Hinblick auf die Endosymbiontentheorie näher untersucht. Obwohl die Interaktion von Proteinen mit Chaperonen eine korrekte Faltung und somit auch die Funktion der Proteine sicherstellt, könnte eine zwingende Anpassung an diese Unterstützung irreversibel sein, was eine evolutionäre Sackgasse darstellt. Korrelationen zwischen Chaperonabhängigkeit und dem Expressionslevel von Proteinen sprechen dafür, dass die Evolution der Chaperon-Substrat Interaktionen maßgeblich von der benötigten Substratmenge begrenzt wird. Der positive Effekt der Chaperon-unterstützten Proteinfaltung auf die organismische *„fitness“* im Sinne des Darwin’schen Evolutionsmodells wird von einem empfindlichen Gleichgewicht zwischen der Abschwächung der negativen Folgen von Fehlfaltungen auf der Einen und einem höheren Energieaufwand sowie einer langsameren Proteinreifung auf der anderen Seite bestimmt.

3 Introduction

Molecular chaperones were first described as proteins that assist in the assembly of other proteins into their functional conformation[1,2]. Besides the assembly of protein complexes and *de novo* folding of nascent polypeptides, chaperones play a role in protein translocation across membranes[3], stabilization of protein-protein interactions[4,5] and ribosome biogenesis[6]. But regardless of their exact function, different chaperones provide assistance in the same assignment: proteins have to maintain their designated function in the right place at the right time.

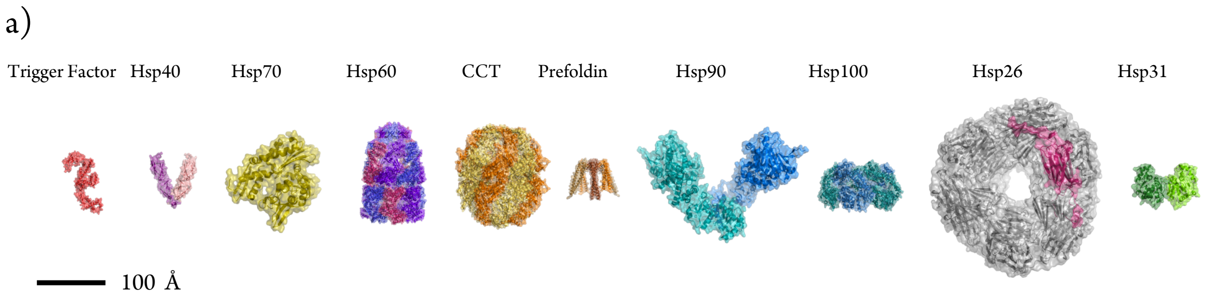
3.1 Chaperone-mediated protein folding in the three domains of life

Species in the three domains of life – eubacteria, archaebacteria and eukaryotes – utilize slightly different chaperones that assemble into diverse protein folding pathways. The major chaperone families in eubacteria are Trigger Factor (TF), DnaJ (Hsp40), DnaK (Hsp70), GrpE (Nucleotide exchange factor), and GroEL/GroES (Hsp60/Hsp10). Trigger Factor is the first chaperone that binds to the nascent polypeptide chain emerging from the ribosome, and its function is to shield hydrophobic (especially aromatic) stretches of the translated protein in order to keep it soluble[7] (Figure 1). Members of the DnaK and DnaJ chaperone families assist protein folding by forming a complex with their substrate proteins. Substrate binding specificity of the ATPase-like DnaK chaperone is determined by the DnaJ co-chaperone[48,49]. Experimental data shows that DnaJ in *E. coli* binds to hydrophobic protein surfaces and initiates the functional cycle of DnaK system by targeting the DnaK to hydrophobic patches within the substrate[50]. DnaK then stabilizes the intermediate conformational state of the substrate using ATP. The nucleotide exchange factor GrpE is involved in binding and release of ATP and ADP. Chaperonin systems comprise barrel-like structures that assists protein folding by providing an isolated environment for the protein to fold[51]. GroE is a eubacterial chaperonin complex composed of two proteins: GroEL, a barrel like structure consisting of two heptameric rings (Figure 1) and GroES, also a heptameric ring, that functions as a lid for the GroEL barrel.

Archaeal species utilize chaperones of the Hsp70[52], Hsp40[53], GrpE[54], and TriC/CCT[55] families. Interestingly, almost all thermophilic species are lacking DnaK and its co-chaperone DnaJ[56]. It is yet unclear if archaea rely on other proteins than DnaK-DnaJ to remove the cellular debris caused by heat shock or if they rely on proteases instead[56]. The existence of a nascent chain-associated complex in Archaebacteria has been experimentally confirmed and it was found to be associated with a ribosome like eukaryotic NAC homolog[57]. The archaeal chaperonin system is termed thermosomes. It forms an octameric double-ring structure with an apical loop instead of a capping cofactor like GroES[58]. *Methanosarcina mazei* is an exception among archaea as it also encodes homologs of eubacterial GroEL and HtpG (Hsp90), which were acquired via a horizontal gene transfer[59]. The substrate set of the two chaperonin systems in *M. mazei* largely overlap, yet the GroE substrates are biased towards proteins with complex α/β domains while the substrates of the thermosome includes a wider range of different domain folds[60]. On the other hand, several eubacterial species, including Clostridial and Cyanobacterial representatives, encode a CCT-like chaperonine[61]. This chaperonine forms a structure that is similar to that of the archaeal CCT and it is thought to be acquired by an ancient horizontal gene transfer during Firmicute evolution[62]. A survey for chaperones in archaeal genome sequences leads to the interesting finding that Hsp90 and Hsp100 are absent in nearly all species[56].

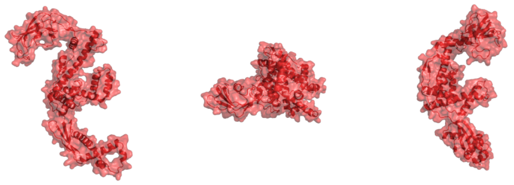
The eukaryotic chaperone repertoire reflects the hybrid origin of the eukaryotic cell. According to the symbiogenic model, eukaryotes evolved from endosymbiosis of two distantly related prokaryotes[63-67]. Later it was substantiated to be the result of an event where a eubacterium, most likely an α- proteobacterium[68-70], was engulfed within an archaebacterial host[71-73]. The eubacterium gave rise to the mitochondrion organelle and the holobiont became what we know today as eukaryotic cells[68]. Subsequently, most of the endosymbiont genomic material was either lost or transferred to the host nucleus, a process that eventually led to the evolution of a mitochondrial protein import apparatus as well as drastic miniaturization of the mitochondrial genome[74-76]. Following this genome reorganization, the complexity of the nuclear genome dramatically increased[77]. Plastids of photosynthetic eukaryotes originated via a similar evolutionary event in which a cyanobacterial endosymbiont was engulfed within a eukaryotic host (see[78] for review). Eukaryotic proteomes are thus mosaics of archaebacterial and eubacterial homologues representing the contribution of the host and organellar ancestors to eukaryotic evolution[79].

Chaperones comprising the eukaryotic protein folding pathway are no exception to that rule and they include homologs to both eubacterial and archaeal chaperones[58,80]. The ribosome-associated complex (RAC) is the first chaperone-complex that interacts with most newly synthesized proteins in *S. cerevisieae*[17]*.* It consists of the HSP40-chaperone Zuo1 and the HSP70 partner Ssz1[81-83]. Further folding of completely translated peptides can be assisted by other HSP40-HSP70 complexes, as well as the HSP90 system and TRiC/CCT class chaperones and their prefoldin co-chaperones[80]. Prefoldin operates mainly on cytoskeleton associated substrate proteins, and assists in their targeting to TRiC/CCT for folding[28]. The TRiC/CCT is a chaperonin system consisting of two rings, but differently from GroE, each ring is formed by eight subunits[84]. This hexadecameric barrel structure is the same as in the archaeal thermosomes, comprising the group of type II chaperonins as opposed to the group I chaperonins GroEL in eubacteria or the mitochondrial Hsp60 (see[85] for review). The eukaryotic TriC/CCT consist of eight different subunits[84], whereas the archaeal thermosome is composed of only two types, the α and β type subunits[86]. Another difference between TRiC/CCT and GroE is that CCT is found to interact with nascent peptide chains more frequently[87] and does not utilize a capping cofactor for the ATP-dependent, GroE-like enclosure and folding process[86]. Nevertheless both chaperonin types are assumed to share similar substrate recognition: GroE and TriC type chaperonins have a substrate overlap of 80% when presented with denatured protein extract of human fibroblasts[88]. The eukaryotic organelles – mitochondria and chloroplast – utilize Hsp60 and Hsp10 chaperones. These are homologs of the eubacterial GroEL and GroES chaperones[89]. The eukaryotic Hsp60 has a “double doughnut” structure similar to GroEL and its expression in the mitochondrion is increased under heat-stress conditions[20]. Hsp60 chaperone interacts with Hsp10 which serves as a capping protein similarly to eubacterial GroES[90]



b)

Trigger Factor



☒ Eubacteria

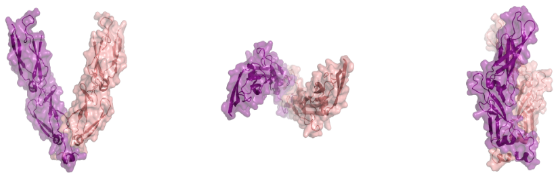
□ Archaea

□ Eukaryotes

Monomer
97 kDa
PDB: 1W26

The eubacterial trigger factor (TF) is associated to nascent polypeptides emerging from the ribosome (8). It projects the extended domains over the exit of the ribosomal tunnel, creating a protected folding space where nascent polypeptides may be shielded from proteases and aggregation (9).

Hsp40 (synonym: DnaJ)



☒ Eubacteria

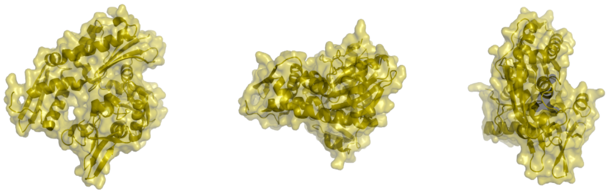
☒ Archaea

☒ Eukaryotes

Dimer
2 x 19 kDa
PDB: 1C3G

Hsp40 is a U-shaped Dimer (10). It functions as a Co-chaperone that stimulates the ATPase activity of the HSP70 chaperone (11). It is involved in protein translocation and the proteolysis of misfolded proteins (12-14).

Hsp70 (synonym: DnaK)



☒ Eubacteria

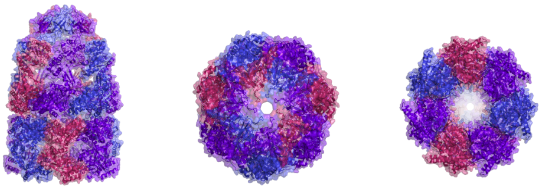
☒ Archaea

☒ Eukaryotes

Monomer
78 kDa
PDB: 3QFP

Hsp70 is a cytoplasmic ATPase (15). It is a ribosome-associated molecular chaperone that is involved in folding of newly made polypeptide chains (16). It functions with a J-protein (Hsp40) partner in the ribosome-associated complex (RAC), which is involved in translocation of proteins into mitochondria as well (17).

Hsp60 (synonyms: GroEL, GroE, Cpn60, mtHsp60)



☒ Eubacteria

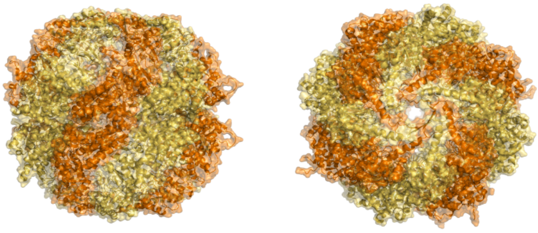
☒ Archaea

☒ Eukaryotes

Tetradecamer
14 x 57 kDa
PDB: 1AON

GroEL is a eubacterial chaperonin that prevents aggregation after heat shock and is required for ATP-dependent folding of precursor polypeptides and complex assembly (18). It has a barrel like structure consisting of 2 heptameric rings composed of 14 identical subunits (19). In Eukaryotes, Hsp60 is localized in the mitochondria and is involved in both protein import (20) and mtDNA transmission (21). GroEL/Hsp60 functions with a capping cofactor termed GroES/Hsp10, which is a ring consisting of seven identical subunits (22) and encapsulates the substrate inside the GroEL molecule (23).

CCT (Synonyms: TriC, Tcp-1, Thermosome)



□ Eubacteria

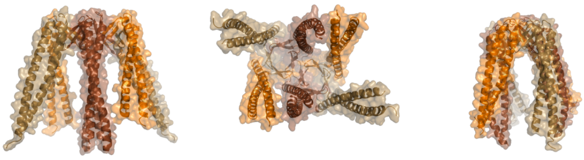
☒ Archaea

☒ Eukaryotes

Hexadecamer
16 x 57..64 kDa
PDB: 3P9D

CCT is the archaeal chaperonin type and is present also in the eukaryotic cytosol. It consists of 2 octameric rings, with 8 or 9 subunits encoded by different genes (24). Unlike GroEL, CCT functions without a cofactor, using a loop in the apical domain to encapsulate substrate proteins (25). CCT is involved in the folding of actin (26) and tubulin (27).

Prefoldin (Synonym: PFD)



□ Eubacteria

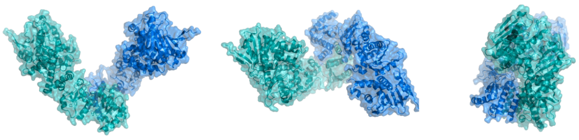
☒ Archaea

☒ Eukaryotes

Heterohexamer
6 x 14..23 kDa
PDB: 1FXK

Prefoldin (PFD) is a heterohexameric chaperone. The archaeal homolog is composed of two subunits, while the eukaryotic PFD is composed of six subunits (28). Prefoldin is present in eukaryotes and archaea, where it binds specifically to cytosolic chaperonin (CCT) and transfers target proteins to it (28).

Hsp90 (synonym: HtpG)



☒ Eubacteria

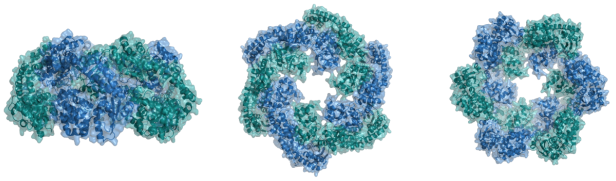
□ Archaea

☒ Eukaryotes

Dimer
2 x 71 kDa (E. coli)
2 x 82 kDa (Yeast)
PDB: 2IOQ

Hsp90 is a dimer that binds its substrates like a molecular clamp (29). It is present in eukaryotes and eubacteria where it is termed htpG (30). Hsp90 is essential in yeast (31). Besides protein folding, it is required for pheromone signaling (32) and preprotein delivery to Tom70p and subsequent translocation into mitochondria (33). Hsp90 also promotes telomerase DNA binding and nucleotide addition (34).

Hsp100 (synonym: Clp)



☒ Eubacteria

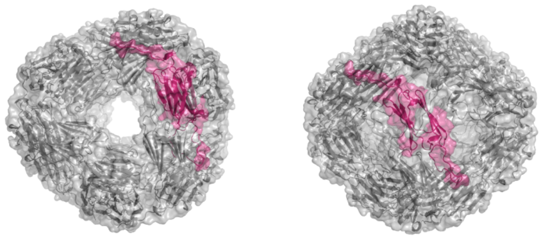
□ Archaea

☒ Eukaryotes

Hexamer
6 x 96 kDa
PDB: 3PXG

Hsp100 is a heat shock protein that refolds and reactivates previously denatured aggregated proteins (35). It cooperates with Ydj1p (Hsp40) and Ssa1p (Hsp70) (36). Homologs of Hsp100 are present in eubacteria, eukaryotes and Mitochondria (37-39). Structurally it is a two-tiered hexamer (40).

Hsp26 (synonym: Hsp16.5)



□ Eubacteria

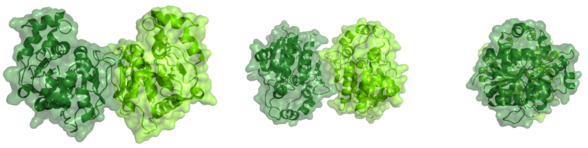
☒ Archaea

☒ Eukaryotes

24mer
24 x 16.5 kDa (Archaea)
24 x 23.7 kDa (Yeast)
PDB: 1SHS

Hsp26 forms hollow oligomers that suppress unfolded protein aggregation. The oligomer activation requires heat induced conformational change. Hsp26 oligomers dissociate into dimers under heat stress. Each dimer binds a substrate monomer. After the substrate binding the dimers are newly assembled forming an Hsp-substrate complex. Hsp26 also has mRNA binding activity. The archaeal homolog is hsp16.5 (41).

Hsp31



☒ Eubacteria

☒ Archaea

☒ Eukaryotes

Dimer
2 x 26.5 kDa
PDB: 1R7W

Hsp31 is a dimeric chaperone and cysteine protease in eukaryotes and bacteria (42-44). It is a member of the DJ-1/ThiJ/PfpI superfamily (45), which includes the human DJ-1 protein that is involved in Parkinson's disease (46). The archaeal homolog is PfpI (47).

Figure 1. Chaperone structural properties. a) Relative complex size of chaperones. b) Chaperone structures are shown in upright orientation (left), in 90° rotation along the Y- (middle) and 90° rotation along X-axis (right). Complexes having a radial symmetry are rotated in 90° along the Y-axis to show the top (middle) and bottom (right) of the molecule. Chaperone plots were generated using PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

3.2 Chaperone-mediated protein unfolding

Molecular chaperones are functional also in unfolding and refolding of previously misfolded proteins[36]. Protein synthesis is energetically the most expensive process within living cells. In bacteria it was estimated that about 60% of the ATP molecules required for the formation of a whole cell are consumed by protein translation[91]. A recent study aiming at quantifying gene expression control in mammalian cells suggested that protein synthesis consumes about 90% of the energy that is needed to maintain the cellular protein levels and determined that protein translation is the limiting factor in protein production[92]. Proteins that fail to fold into their native (functional) state represent an energetic burden of wasted "translation energy". The ability to unfold previously misfolded proteins and reinsert them into the folding pathway is an important process considering the energetic balance of protein production. Protein unfolding/refolding compensates for the fitness cost impaired by the toxicity of protein aggregates in the cell. But maybe even more importantly, the refolding of misfolded proteins means that the energy invested in the synthesis of a misfolded protein will not be wasted. For example, an *in vitro* measurement of the energetic investment in unfolding a Luciferase protein by a DnaK-DnaJ-GrpE complex into its intermediate state revealed that only five ATP molecules are required in this process[93].

Achieving the same outcome by hydrolysis and resynthesis of the hydrogen bonds in Luciferase (550 residues; Swiss-Prot: P08659.1;[94]) is estimated in about 3,000 ATP molecules, hence the rescue of this protein by the chaperones is three orders of magnitude energetically cheaper than its recycling[93].

3.3 Chaperone-mediated protein translocation

In addition to providing co-translational folding mechanisms, molecular chaperones are also involved in protein translocation across membranes, by assisting in stabilizing transported proteins. For example the HSC70 (Hsp70) chaperone in mammals maintains unfolded mitochondrial proteins soluble on their way to the mitochondrial import receptor Tom70[13]. Cytosolic chaperones of the Hsp70 and Hsp90 families can guide preproteins to the Tom70 import receptors in the outer membrane of mitochondria and induce the import process by binding to Tom70 themselves[33]. Mitochondrial Hsp70 forms a motor complex with Tim44 and Mge1 on the inner membrane to facilitate the movement and unfolding of preprotein domains. Together with Mdj1 (Hsp40), Hsp60 and Hsp10 chaperones are also involved in the refolding of already imported proteins in the mitochondria[95]. A similar protein import mechanism was observed in chloroplasts of plant cells involving Hsp70[96], Tic40[97] and chloroplast Hsp60[98]. Chaperones play an important role also in protein trafficking between neighboring cells. For example, in Arabidopsis thaliana the chaperonin TriC-Cct8 was found to be involved in the translocation of KNOTTED1 (KN1) protein through the plasmodesmata channels[99]. The KN1 is an essential transcription factor for the establishment and maintenance of stem cells[100].

3.4 Protein misfolding and fitness

The folding of translated polypeptides into a functional protein is thought to be determined by intrinsic features of the primary sequence as well as environmental factors within the cell[101]. In most cases the native structure of a protein is the one that is also the most stable thermodynamically[102]. Studies of small proteins (60-100 residues) folding dynamics, which convert from their unfolded to their native (functional) state without the complication of highly populated intermediate states, suggest that a few residue interactions within the sequence form a stable folding nucleus around which the rest of the polypeptide rapidly condenses[103]. Misfolding of proteins or protein structure instability is disadvantageous to the cell not only because the protein function is lacking but also due to the formation of protein aggregates. Misfolded proteins tend to cluster in the cell and form long unbranched, and often twisted, fibers of a few nanometers in diameter. A prominent example are amyloid fibrils[104]. The structural characteristics of proteins involved in the formation of amyloid fibers vary from intact globular proteins to large unstructured polypetides but they all share the same common organization with a core structure of β-sheets whose strands run perpendicular to the fibril axis[105]. The formation of misfolded protein aggregates is known to hinder the cell viability. For example, both Alzheimer and Parkinson syndromes are founded in the deposition of protein aggregates in neuronal tissues[106,107].

A recent study[108] quantified the impact of misfolded proteins on the organism fitness by expressing different variants of structurally destabilized yellow-florescent protein (YFP) in yeast cells and measuring their growth rate. The results revealed that an induction of a small amount of YFP aggregates leads to a significant reduction in growth rate. Since the YFP is a gratuitous protein – i.e., its function is not essential in yeast – this result indicates that the presence of protein aggregates alone, regardless of the protein function, imposes a selective cost on the organism.

If protein misfolding imposes selective cost, could it be that the effects of this selective pressure are imprinted in genomes? It has been long known that protein expression level, codon usage, and evolutionary rate are correlated (see Box 1 for an explanation of terms). For almost every sequenced genome tested so far, the proportion of optimal codons within a protein-coding gene is negatively correlated with amino acid substitution rate[109-115]. An analysis of protein abundance in model organisms revealed a significant positive correlation between protein expression level and codon adaptation both in *E. coli*[116] and yeast[111] (Figure 2). The consistency of this expression-codon adaptation-conservation (*ECC* for short) covarion structure led to the suggestion that there exists a single factor underlying these correlations[118]. It has been suggested that protein network properties[119] or protein essentiality[120] play a key role as determinants of the ECC covarion. However, none of these factors provides a plausible explanation for the correlation between protein expression level and the selection against synonymous nucleotide substitutions. Based on the comparison of yeast paralogs having similar protein sequence but different expression levels Drummond et al.[113] concluded that protein expression level is the major determinant of the ECC covarion and suggested that the selection at the DNA level acts against ribosome infidelity during translation in order to minimize protein misfolding. According to their model, the selection against synonymous nucleotide substitutions maintains protein translation accuracy while the selection against non-synonymous nucleotide substitutions maintains the translation robustness[113]. Thus in a model where the fitness is determined by protein translation efficiency, the ECC covarion is determined by selection against mis-translation induced protein misfolding[115].

A strong impact of protein mis-translation on protein folding robustness is extremely important in *in vitro* systems (e.g. Bloom et al.[121]). In living cells, however, misfolding of mis-translated proteins can be compensated for by chaperones. Molecular chaperones lower the energetic barrier for a stable conformation, thus enabling polypeptides that contain destabilizing residues to fold into a functional protein. Chaperone expression level is increased in the presence of unfolded polypeptides, regardless of the type of intracellular and/or environmental stress condition[122,123]. This mechanism of action was exemplified in the recent work by, Geiler-Samerotte et al.[108] which observed an elevation of chaperones expression level – including members of the Hsp70, Hsp40 and Hsp90 families – in the presence of misfolded YFP aggregates. The transcription of many of these chaperones in yeast is activated by the heat shock transcription factor 1 (HSF1) whose availability is conditioned by the presence of misfolded proteins in a negative feedback loop[124]. The role of molecular chaperones in the stress response (especially to heat-stress) has long been studied[125]. By providing proper folding of translated proteins chaperones mitigate the fitness decrease caused by stress induced protein misfolding.

Box 1: Molecular Evolution terms

Gene expression: The process by which the information from a protein-coding or an RNA-specifying gene is used in the synthesis of a gene product. Gene expression is usually measured on a genome-wide scale by using DNA microarrays, which measure the relative amounts of RNA transcripts from thousands of genes at once.

Protein expression: The translation of an mRNA into a protein. Protein expression levels are measured with such technologies as antibody arrays that target specific known proteins or liquid chromatography associated to tandem mass spectrometry. The units of measurement are usually number of protein molecules per cell.

Preferred codon: The most frequently used codon for a particular amino acid.

Codon usage bias: The degree with which codon usage in a protein-coding gene deviates from equal frequencies of occurrence of synonymous codons.

Codon adaptation index (***CAI***): A measure quantifying the deviation of actual codon usage from the optimal codon usage. An optimal codon is the one, among several that encode for the same amino acid, having the highest concentration of corresponding t-RNA in the cell. CAI is calculated assuming that due to selection the most abundant codon for each amino acid is the optimal one[109].

Protein conservation: The degree of similarity between homologous proteins.

Amino-acid replacement rate: The number of amino acid replacements per site per unit time. In comparative studies, the unit time is the divergence time between the two sequences.

Protein connectivity: The number of links that a protein node has to other nodes in the protein-protein interaction network.

Synonymous substitution: A substitution of a nucleotide in the reading frame of a protein-coding gene that results in a change from one codon to a synonymous one. A synonymous substitution does not alter the amino acid encoded by the codon, unless the substitution affects a splicing site or an RNA editing site.

Experimental evolution: Propagating organisms under controlled conditions with the objective to study phenotypic and genotypic changes over time (for review see[126]).

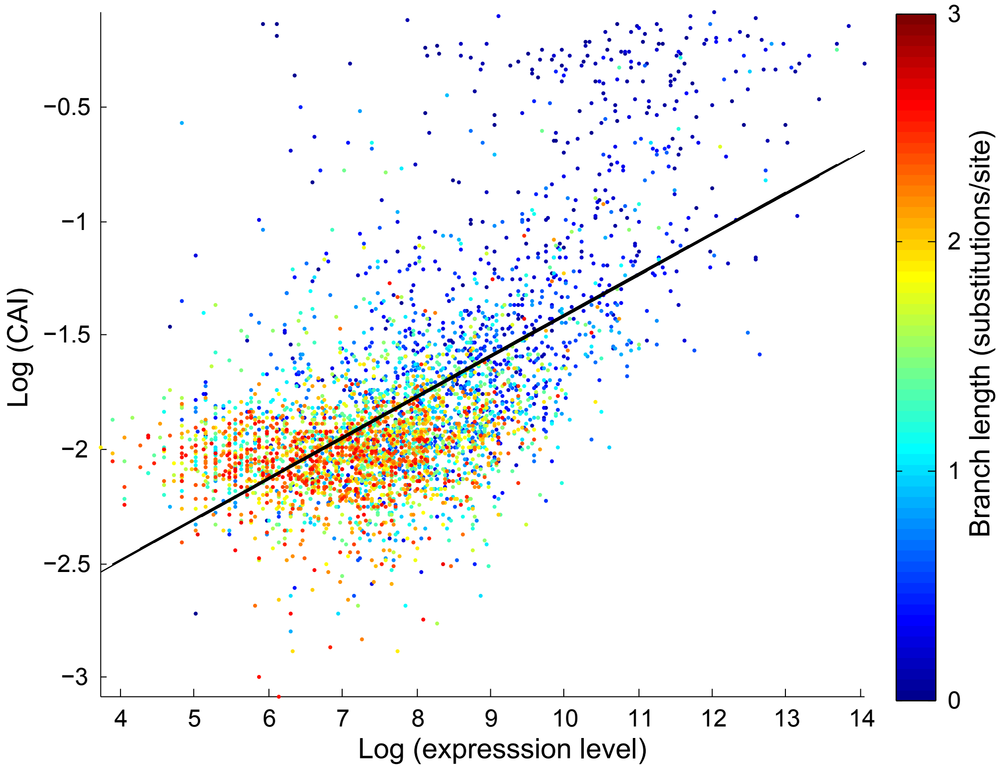


Figure 2. The 3-way correlation between expression level, codon adaptation index (CAI), and evolutionary rate calculated for yeast (data from ([117])).

3.5 Chaperone-mediated protein folding and fitness

Experiments in which chaperone activity in whole organisms was repressed highlighted the extent to which living cells depend upon chaperone-mediated protein folding under normal conditions. A decrease in Hsp90 activity in *Drosophila* by crossing-over with a weak Hsp90 allele (*Hsp83*) or by feeding the flies with an inhibitor of Hsp90 revealed phenotypic deformities that were much more abundant than expected by chance[127]. Applying an inhibitor of Hsp90 activity to *Arabidopsis thaliana* seedlings revealed phenotypes similar to those observed under heat-stress conditions[128]. The resulting phenotypic deformities in these experiments are attributed to the misfolding of Hsp90-clients, many of them are involved in signal transduction[127,128]. The Hsp90 chaperone inhibition in these experiments revealed phenotypic variation that was encoded in the genome but masked by the chaperone activity. This leads to the conclusion that chaperones have a significant impact on the organism's fitness as buffers of phenotypic variation[128,129]. In other words, some genetic variation in protein coding genes has a negligible effect on the phenotype (i.e., it is neutral) as long as the protein conformation – and consequently its function – is maintained constant by the crucial assistance of chaperones[128,129].

Populations facing high mutational loads are prone to suffer from reduced fitness due to destabilizing mutations in protein folding genes leading to protein misfolding. The implication of chaperones as mediators of phenotypic stability suggests that they might be useful for survival in such conditions[129-131]. Indeed, experimental studies of *Escherichia coli* and *Salmonella typhimurium* populations that have been exposed to random mutagenesis revealed that an overexpression of the GroE-chaperonin complex restored[131] or improved[132] their fitness. Fares et al.[131] demonstrated the buffering effect of GroE chaperonin by using a mutator *E. coli* strain[133] that accumulates mutations 3.3-fold faster in comparison to the wild type. After 3,240 generations of random mutation accumulation the fitness (measured by growth rate) of the mutated strains decreased in 50% compared to the ancestral line. Cloning a constitutive GroE operon into the mutated strain resulted on average in 86-fold higher levels of the chaperonin and led to a restored fitness that was only 20% lower in comparison to the ancestral strain. The improved fitness was however conditioned by supplementing the growth media with ample amino acids that were probably required for the GroE translation in massive quantities[131]. This result demonstrates that chaperone overexpression is useful in overcoming high mutational loads, yet there exists a tradeoff between the beneficial impact of the chaperones and the resources required for their production.

A later study by Maisnier-Patin et al.[132] showed that a modest increase in the GroEL expression level is sufficient to improve the fitness. Mutagenesis in *S. typhimurium* was induced by expressing an error-prone DNA-polymerase at different levels, and thaccumulation of random mutations led to a decreased fitness. Samples under high mutational load showed increased expression of the DnaK and GroEL chaperones at a level 2-3 fold higher in comparison to the ancestral strain. The chaperones were probably up-regulated due to the presence of misfolded proteins resulting from the accumulation of destabilizing mutations. Furthermore, an artificial induction of GroEL expression by a factor of ~1.5 improved the fitness substantially[132]. These results supply further evidence that chaperones contribute to antagonistic epistasis where the cumulative effect of mutations in the genome is mitigated[132].

Natural populations evolving under high mutational loads supply further evidence for the buffering effect of chaperones. Microbial endosymbionts are characterized by small population size and effectively no recombination, leading to an increased rate of fixation of deleterious mutations in their genes[134]. Measurements of GroEL concentrations in the bacterium *Buchnera aphidicola* – an intracellular endosymbiont of aphids – showed that it is expressed at a level 7.5-fold higher in comparison to an *E. coli* under normal conditions[135]. This naturally induced overexpression of the chaperonin probably evolved as a compensatory mechanism in order to maintain protein stability under high mutational loads[130,134].

An analysis of the chaperone repertoire in eukaryotic endosymbionts supplies further evidence for the importance of chaperones during reductive evolution. Microsporidia are unicellular eukaryotes, a sister group of fungi, that evolved into obligate intracellular parasites infecting most eukaryotic phyla[136]. Members of the group are characterized by highly reduced genomes encoding very few genes. The number of open reading frames (ORFs) in currently sequenced microsporidal genomes ranges between 1,997 in *Encephalitozoon cuniculi*[137] and 2,633 ORFs in *Trachipleistophora hominis*[138]. A comparison of the microsporida chaperone repertoire to that of yeast reveals an extreme reduction of the Hsp40 and Hsp70 protein families, while all eight genes encoding for the TriC/CCT subunits have been retained. This may suggest that the CCT/TriC chaperones have an essential role in maintaining eukaryotic protein stability under high mutational loads that are typical in reductive genome evolution[139].

3.6 Chaperone-mediated folding and protein evolution

The observations that chaperone expression under high mutational loads can restore or improve the organism's fitness led to the suggestion that protein interaction with chaperones enlarges the spectrum of neutral mutations and consequently increases protein evolvability[129,140]. Using an experimental evolution approach, Tokuriki and Tawfik[141] examined the impact of GroE mediated folding on protein evolution. Various enzymes whose folding (i.e., function) depends upon the GroE chaperonin were exposed to random mutagenesis using an error-prone PCR, and the resulting variants were selected for a further mutagenesis round according to their enzymatic activity. The experiment was performed both under normal conditions and in the presence of overexpressed GroE. The results revealed that overexpression of GroE facilitated the accumulation of significantly more mutations in comparison to the normal mutational drift, and led to the conclusion that protein interaction with the chaperones indeed promotes enzyme evolution.

The finding that GroE increases protein evolvability has been evaluated in an experimental setting. If chaperone-mediate protein evolution occurs also in nature, we might be able to find evidence for it in sequenced genomes. In order to test this hypothesis, one has to compare the evolutionary dynamics of proteins whose folding is assisted by chaperones with proteins that fold independently of the chaperones. Proteins that interact with GroE in *E. coli* can be divided into three classes based on their dependency upon the GroE for folding[142]: GroE-independent proteins (Class I) fold spontaneously in standard conditions (37°C) and attain on average 55% of their activity independent of chaperones, GroE or otherwise. GroE partially dependent proteins (Class II) require GroEL assistance, in addition to other chaperons, at 37°C but do not require GroES at 25°C, where spontaneous folding is observed. GroE obligatory proteins (Class III) fail to fold spontaneously at 37°C and have an obligate requirement for GroE in order to attain activity[142,143]. A comparison of *E. coli* proteins to their orthologs in 446 proteobacterial genomes revealed that obligatory substrates of GroE (Class III) evolve 35% faster than GroE-independent substrates (Class I)[144,145]. The significant difference in amino acids substitution rate among the three

GroE-dependency classes could not be explained by other correlates of protein evolutionary rates such as expression level, protein essentiality, or the number of interactions with other proteins (protein centrality)[144]. These results suggest that during evolution, GroE-mediated folding increases the evolutionary rate of substrate proteins by buffering the deleterious effects of misfolding-related mutations[144,145].

A comparison of codon usage across the three GroE-dependency classes revealed that casual GroE-substrates (Class I) exhibit a higher level of codon- and tRNA adaptation than obligate GroE-substrates (Class III)[144,146,147]. Constraining the comparison of codon usage to buried sites only, which are considered to be structurally sensitive, revealed that the enrichment in optimal codon usage within casual GroE-substrates is even more pronounced[146]. The optimal codon enrichment within the coding sequences of casual GroE-substrates indicates that they are less prone to mistranslation-induced misfolding[146], which fits well with their reduced dependency upon the GroE for folding. Because codon usage and protein expression level are positively correlated, the difference in codon adaptation among the GroE-dependency classes means that casual GroE-substrates are predicted to be more highly expressed in comparison to obligatory substrates[146]. A comparison of protein expression levels measured in *E. coli* strain K12 MG1655[148] among the GroE-dependency classes revealed that this is indeed the case [144].

However, proteins that depend on chaperones for folding have also different physiochemical properties according to the chaperones with which they interact and the degree of their dependency. In an analysis of the impact of chaperone buffering capacity on genome evolution in *E. coli*, strictly dependent substrates of GroE were found to be enriched in positively charged amino acids and in cysteine and proline, and their genes were found to have higher GC content. In addition, the number of protein-protein interactions decreased with the dependency upon GroE[144]. Similarly, protein solubility experiments revealed enriched levels of glycine and alanine in proteins that belong to the most strictly GroE dependent substrate class in *E. coli*[143]. These proteins are also characterized by inherent aggregation propensities that were significantly higher than those of proteins less dependent on GroE for folding.

GroE dependence also correlates with patterns of protein interactions. Casual GroE interactors (Class I) have more protein interactions[144] and are more central in the *E. coli* metabolic network[149] in comparison to obligatory substrates. Hence, proteins that depend upon GroE for folding are found in the periphery of the protein-protein interaction network and the metabolic network[144,149]. These observations led to the suggestion that protein interaction with GroE facilitates the expansion of the metabolic network by enabling substrate proteins to explore their conformation space and evolve novel functions[144,149].

Studying the correlation between chaperone-mediated folding and protein evolution in eukaryotes is complicated by the wealth of chaperones encoded in eukaryotic genomes and the many different folding pathways in which they interact with substrate proteins. A recent large-scale survey of chaperone interactors in *Saccharomyces cerevisiae* using TAP-tag approach revealed that about 60% of the yeast proteome interacts with one or more chaperones[90]. The number of chaperones interacting with a single protein can reach a total of 25 as in the example of Hca4, a putative nucleolar DEAD box RNA helicase. Many chaperones overlap in their subsets of interacting proteins. For example, 63% of the proteins that interact with Ssb1 (Hsp70), interact also with its paralog Ssa1[90]. On the other hand, some chaperones, especially those of the Hsp70 family, can interact with a multitude of substrate proteins, with Sse1 (Hsp70) having the highest number of interacting proteins (2,705 of the 5,880 proteins encoded in yeast)[90]. The global chaperone-protein interaction pattern revealed a positive correlation between the number of interacting chaperones per substrate-protein and the number of hydrophobic stretches in the protein sequence, suggesting the frequency of hydrophobic regions as the phenotypic signal of structurally vulnerable proteins[90].

Within the cytosolic chaperone repertoire the TriC/CCT chaperonin complex was found to have a significant substrate overlap with the eubacterial GroE complex[60,88]. This raises the question whether TriC/CCT influences protein evolution in a similar way to GroE. Warnecke and Hurst[146] searched for detectable evidence for the evolutionary impact of TRiC/CCT in substrate protein sequences. They found proteins that interact with TRiC/CCT to be longer than proteins that do not interact with that chaperone[146]. Yet no correlation between protein expression level and CCT interaction could be observed, despite the fact that longer genes encode less abundant proteins[146]. However, the large substrate overlap and complex interaction patterns in the eukaryotic chaperone interactions network are likely to mask the effect of any single chaperone.

A recent examination of the yeast chaperone-substrate interaction patterns using tools from the field of networks science revealed a remarkable order in the complex chaperone interactions network[117]. An application of modularity function[150] that seeks to divide the network into the most connected components (termed also *communities*) revealed ten communities of proteins and their dedicated chaperones. Five Hsp70 chaperones were not grouped into any community; those interact with more than 1,000 proteins each, and 3,275 proteins in total[90]5, indicating a low substrate specificity in their interaction. Substrate proteins in the ten communities were found to be significantly different in their physiochemical properties such as protein length, the proportion of negative and polar amino acids, aromaticity and the proportion of alpha-helix and coiled-coil secondary structures[117]. Proteins with more chaperone interactions in yeast are longer, heavier and enriched in Aspartate, Glutamate and Lysine amino acids[90]. Proteins with fewer chaperone interactions were found to exhibit higher aromaticity and hydrophobicity and were enriched in Cysteine and Phenylalanine[90]. However the number of hydrophobic stretches of length between one and five was increasing with the number of chaperone interactions. Substrates of the chaperonin TRiC/CCT are enriched in beta-sheets[88]. Proteins with high beta-sheet content were found to be slow folding and vulnerable to misfolding and aggregation[88]. In the network analysis of chaperone-protein interactions in yeast, the substrate proteins in the modules were found to be significantly different not only in the above mentioned biochemical properties but also in the usage of many single amino acids. Aspartate, Glutamate, Glycine, Isoleucine, Leucine, Phenylalanine, Proline and Valine usage was significantly different among the ten modules after a false discovery rate test for multiple comparisons[117].

Using networks approach to analyze the yeast chaperone-substrate interactions network revealed that proteins that interact with different sets of chaperones, are significantly different in their expression level, codon adaptation and sequence conservation. Ranking the chaperone-substrate communities by these three properties shows that they are inter-correlated similarly to the ECC covarion observed in whole genomes[117]. Communities of proteins that are highly expressed are also the communities that evolve with the slowest substitution rates and are encoded by high proportion of preferred codons. Conversely, communities of proteins that have the lowest expression level also evolve in the highest substitution rate and show decreased codon adaptation. Much of the variability in protein substitution rates among the communities is explained by protein expression level, signifying protein abundance within the cell as a major determinant in the ECC covarion.

Chaperones from the Hsp70 family are mostly unspecific in their interaction, but many other chaperones, such as the Hsp40 members[151] and Hsp90 system[152], are. The exact mechanism of substrate recognition by the chaperones is not yet fully understood[151,152]. This is a difficult question to tackle because proteins whose functional folding depends upon the chaperones are probably recognized by characteristics of their intermediate, relatively unstable, structure that is difficult to document using the existing techniques for protein structure determination. Nevertheless, the biased amino acid usage and overrepresentation of particular secondary structure elements in substrates of several chaperone families (e.g., GroE and CCT/TriC) suggest that the information underlying substrate recognition is encoded within the protein sequence. Consequently, proteins that interact with similar chaperones are expected to have common features within their primary and secondary structures. Comparative genomics of proteins classified by their interaction with chaperones revealed that those are significantly different not only in their physiochemical properties[88,90,143] but also in their evolutionary properties[117,144]. These studies implicate protein interaction with chaperones as a major force that shapes the genomic landscape during evolution.

3.7 The evolution of protein interaction with chaperones

The impact of chaperone mediated folding on genomic architecture should be placed in an evolutionary context. How can we make sense of protein interaction with chaperones in light of evolution? We suggest that the origin of molecular chaperones and the evolution of their interaction with substrate proteins can be explained by the constructive neutral evolution model[153], which supplies a possible explanation for the origin of complex biological systems while accounting for the lack of advantage from their intermediate stages[154].

Spontaneous protein folding into a stable structure most probably preceded the origin of chaperones. Thus, chaperones evolved in the presence of spontaneously folding proteins to prevent the aggregation of misfolded polypeptides[155] and functioned at their origin more as “holders” than “folders”. At this stage, the novel function supplied by the chaperones was either beneficial or neutral, imposing only the production costs of the chaperones themselves. The folding assistance provided by chaperones doubtless became beneficial under stress conditions leading protein structural destabilization (e.g. heat shock). Prokaryotes have been shown to evolve with increased mutation rates under stress conditions[156,157]; the buffering supplied by the chaperones could be an essential molecular mechanism in this case. Thus, environmental instability must have played an important role in the emergence of chaperones and their fixation during evolution. Chaperones and their interacting proteins co-evolved and some proteins became obligatory dependent on that interaction. Protein adaptation to the folding assistance of distinct chaperones represents an evolutionary trap that is not easily escaped by random mutational process and drift[158]. Hence chaperone-mediated folding allowed for an increased structural complexity at the cost of an obligatory requirement for the chaperones.

The translation of proteins that require the assistance of molecular chaperones for folding has to be coordinated with the chaperone interaction. Recent studies revealed an important role of codon usage and codon usage distribution along the gene in controlling protein translation speed dynamics[159,160]. Casual GroE-substrates in *E. coli*, that can also fold into their functional structure spontaneously, are encoded by a higher proportion of preferred codons in comparison to obligatory substrates and are also more abundant in the cell[144,146], which fits well with the ECC covarion. We suggest that this bias stems from the requirement for synchronization between protein translation and cotranslational folding[161]. Nascent polypeptides that are able to fold spontaneously into their functional conformation are free from that constraint and can be translated at a higher speed. Moreover, it is possible that in order to gain a stable conformation, the whole nascent polypeptide should be available before folding. However, with increasing translation speed the fitness cost of misfolding also increases drastically. Consequently accuracy becomes more important so that proteins that are translated at high speed should also be more conserved[115].

The evolution of protein interaction with chaperones should be inspected also from the systemic point of view. A recent survey for GroEL interactors in *E. coli* revealed that 794 proteins (~18% of the *E. coli* proteome) interact with the chaperonin[162]. Out of the 5,880 proteins in yeast, 595 where found to interact with the CCT/TriC chaperone[90]. We propose that the required protein abundance in the cell largely determines the kind and mode of interaction of that protein with molecular chaperones for folding. The first reason is the energetic investment in chaperone-mediated folding. Chaperone-mediated folding by itself, does not require much ATP in comparison to the translation process. For example, the GroE chaperonin consumes 7 ATP molecules in each round of substrate turnover[163], while translation of a single amino acid costs four ATP molecules[91]. The average protein sequence length in *E. coli* is 316 hence one round within the chaperonin will add only 0.5% to the ATP consumption of the protein production. However, if GroE is required for the folding of many proteins, then GroE in itself should be highly expressed. Moreover, if it is required for the production of highly expressed proteins then it should be produced in even higher quantities. The GroE production costs amount to translation of seven GroES subunits (7x97 amino acids) and 14 GroEL subunits (14x548 amino acids). Apparently the constitutive production of GroE creates an overload of the translation system and an arrest of cell growth[131]. Furthermore, each round of folding by GroE takes about 10 seconds[163], which may slow down protein production considerably. This indicates that chaperone attention should be limited according to the available energetic resources and temporal dynamics of protein synthesis within the cell. Large-scale analysis of chaperone interaction data supports that notion. A comparison of expression level between GroE-dependence groups showed that casual substrates are more highly expressed in comparison to obligatory interactors[144-146]. Similarly, yeast proteins that interact only with one of the promiscuous HSP70 chaperones are more highly expressed in comparison to proteins that interact with additional chaperones[117].

Studies of hemoglobin polymerization *in vitro* showed that polymer formation rate depends on the concentration of soluble monomers[164]. Existing polymers serve as a basis for the formation of heterogeneous polymers. Thus hemoglobin polymerization is an autocatalytic process whose rate is log-linear proportional to the monomer concentration[164]. This idea was recently adopted for the formation of amyloid fibrils[165]. Taken together, these studies indicate that the formation of protein aggregates within the cell largely depend on the abundance of misfolded proteins. This could act as an additional negative selection pressure that keeps highly expressed proteins from developing dependency upon the chaperones for folding because failure in the folding stage will lead to a massive amount of misfolded proteins in a very short time.

In summary, chaperones are crucial in enabling many nascent polypeptides to attain their functional conformation and in providing an energetically efficient mechanism for the recycling of misfolded proteins. Genomic data reveals that chaperones have an important role in shaping genomic landscapes, stemming from the part they play in the intricate correlation between expression level, translation rate, codon usage and sequence conservation. In a broader evolutionary context, molecular chaperones mitigate the deleterious effects of protein misfolding, thus enabling a wider range of genetic variability - the raw material for positive selection, adaptation and innovation.

3.8 Thematic content

This thesis, dealing with the impact of chaperone-mediated protein-folding during evolution, consists of three published research articles.

The first publication deals with the impact of chaperonin-dependent folding on prokaryotic genome evolution. Proteins that are obligatory folded with the assistance of the GroEL/GroES chaperonin complex show increased substitution rates compared to proteins that do not share this strict dependency. Stabilizing protein folding enables chaperones to mitigate the negative outcome of deleterious mutations and thus provide certain proteins with an increased evolvability.

The second part of the thesis describes the community structure in the chaperone interaction network of *S. cerevisiae*. Chaperone interactions define groups of proteins that are characterized by similar expression levels and evolutionary rates. The results indicate that the evolution of chaperone-substrate interactions is bounded by the requirement for protein abundance in the cell.

The third part elucidates the evolutionary history of eukaryotic chaperones in light of the endosymbiosis theory for the origin of eukaryotes. The origin of chaperones in the model organism *Saccharomyces cerevisiae* was traced to one of the two prokaryotic domains, eubacteria and archaebacteria, that were involved in eukaryogenesis. Remarkably *S. cerevisiae* contains nearly the whole chaperone repertoires of eubacteria and archaebacteria. The protein folding machinery of current eukaryotes is a fully integrated system and protein interactions with the chaperones do no longer harbor the signal of its mosaic origin.

This introductory chapter is itself published as a review article: Bogumil D., and Dagan, T. (2012) Cumulative Impact of Chaperone-Mediated Folding on Genome Evolution. Biochemistry 51, 9941-9953.

4 References

(1) Laskey, R., Honda, B., Mills, A., and Finch, J. (1978) Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature* 275, 416-420.

(2) Ellis, J. (1987) Proteins as molecular chaperones. *Nature* 328, 378–379.

(3) Becker, J., Walter, W., Yan, W., and Craig, E. (1996) Functional interaction of cytosolic hsp70 and a DnaJ-related protein, Ydj1p, in protein translocation in vivo. *Mol. Cell Biol.* 16, 4378–4386.

(4) Hartl, F. U. (1996) Molecular chaperones in cellular protein folding. *Nature* 381, 571-580.

(5) Ellis, R. J. (2006) Molecular chaperones: assisting assembly in addition to folding. *Trends Biochem. Sci.* 31, 395–401.

(6) Albanese, V., Reissmann, S., and Frydman, J. (2010) A ribosome-anchored chaperone network that facilitates eukaryotic ribosome biogenesis. *J. Cell Biol.* 189, 69–81.

(7) Patzelt, H., Rüdiger, S., Brehmer, D., Kramer, G., Vorderwülbecke, S., Schaffitzel, E., Waitz, A., Hesterkamp, T., Dong, L., Schneider-Mergener, J., Bukau, B., and Deuerling, E. (2001) Binding specificity of *Escherichia coli* trigger factor. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14244–14249.

(8) Hesterkamp, T., Hauser, S., Lütcke, H., and Bukau, B. (1996) *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* 93, 4437–4441.

(9) Ferbitz, L., Maier, T., Patzelt, H., Bukau, B., Deuerling, E., and Ban, N. (2004) Trigger factor in complex with the ribosome forms a molecular cradle for nascent proteins. *Nature* 431, 590–596.

(10) Sha, B., Lee, S., and Cyr, D. M. (2000) The crystal structure of the peptide-binding fragment from the yeast Hsp40 protein Sis1. *Structure* 8, 799–807.

(11) Cyr, D., Langer, T., and Douglas, M. (1994) DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70. *Trends Biochem. Sci.* 19, 176-181.

(12) Caplan, A. J., Cyr, D., and Douglas, M. (1992) Ydj1p Facilitates Polypeptide Translocation Across Different Intracellular Membranes by a Conserved Mechanism. *Cell* 71, 1143–1155.

(13) Deshaies, R. J., Koch, B. D., Werner-Washburne, M., Craig, E. A., and Schekman, R. (1988) A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. *Nature* 332, 800–805.

(14) Walsh, P., Bursać, D., Law, Y. C., Cyr, D., and Lithgow, T. (2004) The J-protein family: modulating protein assembly, disassembly and translocation. *EMBO Rep.* 5, 567–571.

(15) Zylicz, M., LeBowitz, J. H., McMacken, R., and Georgopoulos, C. (1983) The dnaK protein of *Escherichia coli* possesses an ATPase and autophosphorylating activity and is essential in an in vitro DNA replication system. *Proc. Natl. Acad. Sci. U.S.A.* 80, 6431-6435.

(16) Beckmann, R. P., Mizzen, L., and Welch, W. J. (1990) Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. *Science* 248, 850-854.

(17) Gautschi, M., Lilie, H., Funfschilling, U., Mun, A., Ross, S., Lithgow, T., Rucknagel, P., and Rospert, S. (2001) RAC, a stable ribosome-associated complex in yeast formed by the DnaK-DnaJ homologs Ssz1p and zuotin. *Proc. Natl. Acad. Sci. U.S.A.* 98, 3762–3767.

(18) Thirumalai, D., and Lorimer, G. H. (2001) Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 30, 245–269.

(19) Hendrix, R. W. (1979) Purification and Properties of GroE, a Host Protein Involved in Bacteriophage Assembly. *J. Mol. Biol.* 129, 375–392.

(20) Cheng, M. Y., Hartl, F. U., Martin, J., Pollock, R. A., Kalousek, F., Neupert, W., Hallberg, E. M., Hallberg, R. L., and Horwich, A. L. (1989) Mitochondrial heat-shock protein hsp60 is essential for assembly of proteins imported into yeast mitochondria. *Nature* 337, 620–625.

(21) Kaufman, B. A. (2003) A function for the mitochondrial chaperonin Hsp60 in the structure and transmission of mitochondrial DNA nucleoids in *Saccharomyces cerevisiae*. *J. Cell Biol.* 163, 457–461.

(22) Hunt, J. F., Weaver, A. J., Landry, S. J., Gierasch, L., and Deisenhofer, J. (1996) The crystal structure of the GroES co-chaperonin at 2.8 A resolution. *Nature* 379, 37–45.

(23) Langer, T., Pfeifer, G., Martin, J., Baumeister, W., and Hartl, F. U. (1992) Chaperonin-Mediated Protein Folding - Groes Binds to One End of the Groel Cylinder, Which Accommodates the Protein Substrate Within Its Central Cavity. *Embo J.* 11, 4757–4765.

(24) Kubota, H., Hynes, G., Carne, A., Ashworth, A., and Willison, K. (1994) Identification of six Tcp-1- related genes encoding divergent subunits of the TCP-1-containing chaperonin. *Curr. Biol.* 4, 89–99.

(25) Klumpp, M., Baumeister, W., and Essen, L. O. (1997) Structure of the substrate binding domain of the thermosome, an archaeal group II chaperonin. *Cell* 91, 263–270.

(26) Gao, Y., Thomas, J. O., Chow, R. L., Lee, G. H., and Cowan, N. J. (1992) A cytoplasmic chaperonin that catalyzes beta-actin folding. *Cell* 69, 1043–1050.

(27) Frydman, J., Nimmesgern, E., Erdjument-Bromage, H., Wall, J. S., Tempst, P., and Hartl, F. U. (1992) Function in Protein Folding of Tric, a Cytosolic Ring Complex Containing Tcp-1 and Structurally Related Subunits. *Embo J.* 11, 4767–4778.

(28) Vainberg, I., Lewis, S., Rommelaere, H., Ampe, C., Vandekerckhove, J., Klein, H., and Cowan, N. (1998) Prefoldin, a chaperone that delivers unfolded proteins to cytosolic chaperonin. *Cell* 93, 863–873.

(29) Prodromou, C., Roe, S. M., Piper, P. W., and Pearl, L. H. (1997) A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone. *Nat. Struct. Mol. Biol.* 4, 477–482.

(30) Bardwell, J. C., and Craig, E. A. (1987) Eukaryotic Mr 83,000 heat shock protein has a homologue in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 84, 5177–5181.

(31) Borkovich, K. A., Farrelly, F. W., Finkelstein, D. B., Taulien, J., and Lindquist, S. (1989) hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. *Mol. Cell. Biol.* 9, 3919–3930.

(32) Louvion, J. F., Abbas-Terki, T., and Picard, D. (1998) Hsp90 is required for pheromone signaling in yeast. *Mol. Biol. Cell* 9, 3071–3083.

(33) Young, J., Hoogenraad, N., and Hartl, F. (2003) Molecular chaperones Hsp90 and Hsp70 deliver preproteins to the mitochondrial import receptor Tom70. *Cell* 112, 41–50.

(34) Toogun, O. A., DeZwaan, D. C., and Freeman, B. C. (2007) The Hsp90 Molecular Chaperone Modulates Multiple Telomerase Activities. *Mol. Cell. Biol.* 28, 457–467.

(35) Parsell, D. A., Kowal, A. S., Singer, M. A., and Lindquist, S. (1994) Protein disaggregation mediated by heat-shock protein Hsp104. *Nature* 372, 475–478.

(36) Glover, J., and Lindquist, S. (1998) Hsp104, Hsp70, and Hsp40: A novel chaperone system that rescues previously aggregated proteins. *Cell* 94, 73–82.

(37) Mogk, A., Tomoyasu, T., Goloubinoff, P., Rüdiger, S., Röder, D., Langen, H., and Bukau, B. (1999) Identification of thermolabile *Escherichia coli* proteins: prevention and reversion of aggregation by DnaK and ClpB. *Embo J.*18, 6934–6949.

(38) Queitsch, C., Hong, S. W., Vierling, E., and Lindquist, S. (2000) Heat shock protein 101 plays a crucial role in thermotolerance in Arabidopsis. *Plant Cell* 12, 479–492.

(39) Schmitt, M., Neupert, W., and Langer, T. (1996) The molecular chaperone Hsp78 confers compartment-specific thermotolerance to mitochondria. *J. Cell Biol.* 134, 1375–1386.

(40) Lee, S., Sowa, M. E., Watanabe, Y. H., Sigler, P. B., Chiu, W., Yoshida, M., and Tsai, F. (2003) The structure of clpB: A molecular chaperone that rescues proteins from an aggregated state. *Cell* 115, 229–240.

(41) Haslbeck, M., Walke, S., Stromer, T., Ehrnsperger, M., White, H., Chen, S., Saibil, H., and Buchner, J. (1999) Hsp26: a temperature-regulated chaperone. *Embo J.* 18, 6744–6751.

(42) Quigley, P. M., Korotkov, K., Baneyx, F., and Hol, W. G. J. (2003) The 1.6-Å crystal structure of the class of chaperones represented by *Escherichia coli* Hsp31 reveals a putative catalytic triad. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3137-3147.

(43) Lee, S. J. (2003) Crystal Structures of Human DJ-1 and *Escherichia coli* Hsp31, Which Share an Evolutionarily Conserved Domain. *J. Biol. Chem.* 278, 44552–44559.

(44) Zhao, Y., Liu, D., Kaluarachchi, W. D., Bellamy, H. D., White, M. A., and Fox, R. O. (2003) The crystal structure of *Escherichia coli* heat shock protein YedU reveals three potential catalytic active sites. *Protein Sci.* 12, 2303–2311.

(45) Wilson, M. A., Amour, C. V. S., Collins, J. L., Ringe, D., and Petsko, G. A. (2004) The 1.8-Å resolution crystal structure of YDR533Cp from *Saccharomyces cerevisiae*: A member of the DJ-1/ThiJ/PfpI superfamily. *Proc. Natl. Acad. Sci. U.S.A.* 101, 1531-1536.

(46) Bonifati, V. (2002) Mutations in the DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. *Science* 299, 256–259.

(47) Halio, S. B., Blumentals, I. I., Short, S. A., Merrill, B. M., and Kelly, R. M. (1996) Sequence, expression in *Escherichia coli*, and analysis of the gene encoding a novel intracellular protease (PfpI) from the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* 178, 2605–2612.

(48) Zylicz, M., Ang, D., Liberek, K., and Georgopoulos, C. (1989) Initiation of lambda DNA replication with purified host- and bacteriophage-encoded proteins: the role of the dnaK, dnaJ and grpE heat shock proteins. *Embo J.* 8, 1601–1608.

(49) Hoffmann, H. J., Lyman, S. K., Lu, C., Petit, M. A., and Echols, H. (1992) Activity of the Hsp70 chaperone complex--DnaK, DnaJ, and GrpE--in initiating phage lambda DNA replication by sequestering and releasing lambda P protein. *Proc. Natl. Acad. Sci. U.S.A.* 89, 12108–12111.

(50) Rüdiger, S., Schneider-Mergener, J., and Bukau, B. (2001) Its substrate specificity characterizes the DnaJ co-chaperone as a scanning factor for the DnaK chaperone. *Embo J.* 20, 1042–1050.

(51) Chen, S., Roseman, A. M., Hunter, A. S., Wood, S. P., Burston, S. G., Ranson, N. A., Clarke, A. R., and Saibil, H. R. (1994) Location of a folding protein and shape

changes in GroEL-GroES complexes imaged by cryo-electron microscopy. *Nature* 371, 261–264.

(52) Macario, A. J., Dugan, C. B., and Conway de Macario, E. (1991) A dnaK homolog in the archaebacterium *Methanosarcina mazei* S6. *Gene* 108, 133–137.

(53) Macario, A. J., Dugan, C. B., Clarens, M., and Conway de Macario, E. (1993) dnaJ in Archaea. *Nucleic Acids Res.* 21, 2773.

(54) Conway de Macario, E., Dugan, C. B., and Macario, A. J. (1994) Identification of a grpE heat-shock gene homolog in the archaeon *Methanosarcina mazei*. *J. Mol. Biol.* 240, 95–101.

(55) Rommelaere, H., Van Troys, M., Gao, Y., Melki, R., Cowan, N. J., Vandekerckhove, J., and Ampe, C. (1993) Eukaryotic cytosolic chaperonin contains t-complex polypeptide 1 and seven related subunits. *Proc. Natl. Acad. Sci. U.S.A.* 90, 11975–11979.

(56) Large, A. T., Goldberg, M. D., and Lund, P. A. (2009) Chaperones and protein folding in the archaea. *Biochem. Soc. Trans.* 37, 46-51.

(57) Spreter, T. (2005) The Crystal Structure of Archaeal Nascent Polypeptide-associated Complex (NAC) Reveals a Unique Fold and the Presence of a Ubiquitin-associated Domain. *J. Biol. Chem.* 280, 15849–15854.

(58) Large, A. T., and Lund, P. A. (2009) Archaeal chaperonins. *Front. Biosci.* 14, 1304–1324.

(59) Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R., Martinez-Arias, R., Henne, A., Wiezer, A., Baumer, S., Jacobi, C., Bruggemann, H., Lienard, T., Christmann, A., Bomeke, M., Steckel, S., Bhattacharyya, A., Lykidis, A., Overbeek, R., Klenk, H., Gunsalus, R., Fritz, H., and Gottschalk, G. (2002) The genome of *Methanosarcina maze*i: Evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol.* 4, 453–461.

(60) Hirtreiter, A. M., Calloni, G., Forner, F., Scheibe, B., Puype, M., Vandekerckhove, J., Mann, M., Hartl, F. U., and Hayer-Hartl, M. (2009) Differential substrate specificity of group I and group II chaperonins in the archaeon *Methanosarcina mazei*. *Mol. Microbiol.* 74, 1152–1168.

(61) Williams, T. A., Codoñer, F. M., Toft, C., and Fares, M. A. (2010) Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends Genet.* 26, 47–51.

(62) Techtmann, S. M., and Robb, F. T. (2010) Archaeal-like chaperonins in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20269–20274.

(63) Alvarez-Ponce, D., Lopez, P., Bapteste, E., and McInerney, J. O. (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1594–1603.

(64) Embley, T. M., and Martin, W. (2006) Eukaryotic evolution, changes and challenges. *Nature* 440, 623–630.

(65) Pisani, D., Cotton, J. A., and McInerney, J. O. (2007) Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes. *Mol. Biol. Evol.* 24, 1752–1760.

(66) Rivera, M. C., and Lake, J. A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431, 152–155.

(67) Sagan, L. (1967) On the origin of mitosing cells. *J. Theoret. Biol.* 14, 225–274.

(68) Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., and Leister, D. (2004) A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21, 1643–1660.

(69) Gabaldón, T., and Huynen, M. A. (2003) Reconstruction of the proto-mitochondrial metabolism. *Science* 301, 609–609.

(70) Gray, M. W. (1999) Mitochondrial Evolution. Science 283, 1476–1481.

(71) Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. (2008) The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20356–20361.

(72) Martin, W., and Muller, M. (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392, 37–41.

(73) Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J., and Embley, T. M. (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. *P. Roy. Soc. B-Biol. Sci.* 279, 4870–4879.

(74) Martin, W. (2003) Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8612–8614.

(75) Martin, W., and Herrmann, R. (1998) Gene transfer from Organelles to the Nucleus: How Much, What Happens, and Why? *Plant Physiol.* 118, 9–17.

(76) Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–135.

(77) Lane, N., and Martin, W. (2010) The energetics of genome complexity. *Nature* 467, 929–934.

(78) Gould, S. B., Waller, R. F., and McFadden, G. I. (2008) Plastid Evolution. *Annu. Rev. Plant Biol.* 59, 491–517.

(79) Thiergart, T., Landan, G., Schenk, M., Dagan, T., and Martin, W. F. (2012) An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin. *Genome Biol. Evol.* 4, 466–485.

(80) Young, J., Agashe, V., Siegers, K., and Hartl, F. (2004) Pathways of chaperone-mediated protein folding in the cytosol. *Nat. Rev. Mol. Cell Biol.* 5, 781–791.

(81) Michimoto, T., Aoki, T., Toh-e, A., and Kikuchi, Y. (2000) Yeast Pdr13p and Zuo1p molecular chaperones are new functional Hsp70 and Hsp40 partners. *Gene* 257, 131–137.

(82) Yan, W., Schilke, B., Pfund, C., Walter, W., Kim, S., and Craig, E. (1998) Zuotin, a ribosome-associated DnaJ molecular chaperone. *Embo J.* 17, 4809–4817.

(83) Gautschi, M., Mun, A., Ross, S., and Rospert, S. (2002) A functional chaperone triad on the yeast ribosome. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4209–4214.

(84) Valpuesta, J., Martin-Benito, J., Gomez-Puertas, P., Carrascosa, J., and Willison, K. (2002) Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT. *FEBS Lett.* 529, 11–16.

(85) Lund, P., Large, A., and Kapatai, G. (2003) The chaperonins: perspectives from the Archaea. *Biochem. Soc. Trans.* 31, 681–685.

(86) Ditzel, L., Lowe, J., Stock, D., Stetter, K., Huber, H., Huber, R., and Steinbacher, S. (1998) Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell* 93, 125–138.

(87) McCallum, C., Do, H., Johnson, A., and Frydman, J. (2000) The interaction of the chaperonin tailless complex polypeptide 1 (TCP1) ring complex (TRiC) with ribosome-bound nascent chains examined using photo-cross-linking. *J. Cell Biol.* 149, 591–601.

(88) Yam, A. Y., Xia, Y., Lin, H.-T. J., Burlingame, A., Gerstein, M., and Frydman, J. (2008) Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nat. Struct. Mol. Biol.* 15, 1255–1262.

(89) Hemmingsen, S. M., Woolford, C., van der Vies, S. M., Tilly, K., Dennis, D. T., Georgopoulos, C. P., Hendrix, R. W., and Ellis, R. J. (1988) Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* 333, 330–334.

(90) Gong, Y., Kakihara, Y., Krogan, N., Greenblatt, J., Emili, A., Zhang, Z., and Houry, W. A. (2009) An atlas of chaperone-protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. *Mol. Syst. Biol.* 5, 1–14.

(91) Stouthamer, A. H. (1973) A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie van Leeuwenhoek* 39, 545–565.

(92) Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337–342.

(93) Sharma, S. K., De Los Rios, P., Christen, P., Lustig, A., and Goloubinoff, P. (2010) The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. *Nat. Chem. Biol.* 6, 914–920.

(94) De Wet, J. R., Wood, K. V., DeLuca, M., Helinski, D. R., and Subramani, S. (1987) Firefly luciferase gene: structure and expression in mammalian cells. *Mol. Cell. Biol.* 7, 725–737.

(95) Voos, W., and Rottgers, K. (2002) Molecular chaperones as essential mediators of mitochondrial biogenesis. *Biochim. Biophys. Acta* 1592, 51–62.

(96) Zhang, X., and Glaser, E. (2002) Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 7, 14–21.

(97) Chou, M., Fitzpatrick, L., Tu, S., Budziszewski, G., Potter-Lewis, S., Akita, M., Levin, J., Keegstra, K., and Li, H. (2003) Tic40, a membrane-anchored co-chaperone homolog in the chloroplast protein translocon. *Embo J.* 22, 2970–2980.

(98) Lubben, T. H., Donaldson, G. K., Viitanen, P. V., and Gatenby, A. A. (1989) Several proteins imported into chloroplasts form stable complexes with the GroEL-related chloroplast molecular chaperone. *Plant Cell* 1, 1223–1230.

(99) Xu, X. M., Wang, J., Xuan, Z., Goldshmidt, A., Borrill, P. G. M., Hariharan, N., Kim, J. Y., and Jackson, D. (2011) Chaperonins Facilitate KNOTTED1 Cell-to-Cell Trafficking and Stem Cell Function. *Science* 333, 1141–1144.

(100) Lucas, W. J., Bouché-Pillon, S., Jackson, D. P., Nguyen, L., Baker, L., Ding, B., and Hake, S. (1995) Selective trafficking of KNOTTED1 homeodomain protein and its mRNA through plasmodesmata. *Science* 270, 1980–1983.

(101) Dobson, C. M. (2003) Protein folding and misfolding. *Nature* 426, 884–890.

(102) Dobson, C., Sali, A., and Karplus, M. (1998) Protein folding: A perspective from theory and experiment. *Angew. Chem., Int. Ed.* 37, 868–893.

(103) Fersht, A. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 97, 1525–1529.

(104) Colon, W., and Kelly, J. W. (1992) Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. *Biochemistry* 31, 8654–8660.

(105) Sunde, M., Serpell, L. C., Bartlam, M., Fraser, P. E., Pepys, M. B., and Blake, C. C. (1997) Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* 273, 729–739.

(106) Hardy, J., and Allsop, D. (1991) Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.* 12, 383-388.

(107) Braak, H., and Braak, E. (1990) Cognitive impairment in Parkinson's disease: Amyloid plaques, neurofibrillary tangles, and neuropil threads in the cerebral cortex. *J. Neural Transm.: Parkinson's Dis. Dementia Sect.* 2, 45–57.

(108) Geiler-Samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M., Hartl, D. L., and Drummond, D. A. (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 108, 680-685.

(109) Sharp, P. M., and Li, W. H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295.

(110) Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9, 43–74.

(111) Pál, C., Papp, B., and Hurst, L. D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931.

(112) Krylov, D. M. (2003) Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. *Genome Res.* 13, 2229–2235.

(113) Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14338–14343.

(114) Pál, C., Papp, B., and Lercher, M. J. (2006) An integrated view of protein evolution. *Nat. Rev. Genet.* 7, 337–348.

(115) Drummond, D. A., and Wilke, C. O. (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* 134, 341–352.

(116) Rocha, E. P. C. (2003) An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Mol. Biol. Evol.* 21, 108–116.

(117) Bogumil, D., Landan, G., Ilhan, J., and Dagan, T. (2012) Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate. *Genome Biol. Evol.* 4, 618–625.

(118) Drummond, D. A., Raval, A., and Wilke, C. O. (2006) A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 23, 327–337.

(119) Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752.

(120) Hirsh, A., and Fraser, H. (2001) Protein dispensability and rate of evolution. *Nature* 411, 1046–1049.

(121) Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5869–5874.

(122) Ananthan, J., Goldberg, A. L., and Voellmy, R. (1986) Abnormal proteins serve as eukaryotic stress signals and trigger the activation of heat shock genes. *Science* 232, 522–524.

(123) Baler, R., Welch, W. J., and Voellmy, R. (1992) Heat shock gene regulation by nascent polypeptides and denatured proteins: hsp70 as a potential autoregulatory factor. *J. Cell Biol.* 117, 1151–1159.

(124) Zou, J., Guo, Y., Guettouche, T., Smith, D., and Voellmy, R. (1998) Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. *Cell* 94, 471–480.

(125) Lindquist, S. (1986) The heat-shock response. *Annu. Rev. Biochem.* 55, 1151–1191.

(126) Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012) New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nat. Rev. Microbiol.* 10, 352–365.

(127) Rutherford, S. L., and Lindquist, S. (1998) Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342.

(128) Queitsch, C., Sangster, T., and Lindquist, S. (2002) Hsp90 as a capacitor of phenotypic variation. *Nature* 417, 618–624.

(129) Rutherford, S. L. (2003) Between genotype and phenotype: protein chaperones and evolvability. *Nat. Rev. Genet.* 4, 263–274.

(130) Moran, N. A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 93, 2873.

(131) Fares, M. A., Ruiz-González, M. X., Moya, A., Elena, S. F., and Barrio, E. (2002) GroEL buffers against deleterious mutations. *Nature* 417, 398.

(132) Maisnier-Patin, S., Roth, J. R., Fredriksson, Å., Nyström, T., Berg, O. G., and Andersson, D. I. (2005) Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat. Genet.* 37, 1376–1379.

(133) de Visser, J. A. G. M., Zeyl, C. W., Gerrish, P. J., Blanchard, J. L., and Lenski, R. E. (1999) Diminishing returns from mutation supply rate in asexual populations. *Science* 283, 404-406.

(134) McCutcheon, J. P., and Moran, N. A. (2011) Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26.

(135) Baumann, P., Baumann, L., and Clark, M. (1996) Levels of *Buchnera aphidicola* chaperonin GroEL during growth of the aphid *Schizaphis graminum*. *Curr. Microbiol.* 32, 279–285.

(136) Corradi, N., and Slamovits, C. H. (2011) The intriguing nature of microsporidian genomes. *Briefings Funct. Genomics* 10, 115–124.

(137) Katinka, M. D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., Alaoui, El, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C. P. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453.

(138) Heinz, E., and Lithgow, T. (2012) Back to basics: A revealing secondary reduction of the mitochondrial protein import pathway in diverse intracellular parasites. *Biochim. Biophys. Acta* 1833, 295-303.

(139) Heinz, E., Williams, T. A., Nakjang, S., Noel, C. J., Swan, D. C., Goldberg, A. V., Harris, S. R., Weinmaier, T., Markert, S., Becher, D., Bernhardt, J., Dagan, T., Hacker, C., Schweder, T., Rattei, T., Hirt, R. P., and Embley, T. M. (2012) The

Genome of the Obligate Intracellular Parasite *Trachipleistophora hominis*: New Insights into Microsporidian Genome Dynamics and Reductive Evolution. *Plos Pathog.* 8, e1002979.

(140) Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19, 596–604.

(141) Tokuriki, N., and Tawfik, D. S. (2009) Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* 459, 668–673.

(142) Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H.-C., Stines, A. P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., Mann, M., and Hartl, F. U. (2005) Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in *Escherichia coli. Cell* 122, 209–220.

(143) Fujiwara, K., Ishihama, Y., Nakahigashi, K., Soga, T., and Taguchi, H. (2010) A systematic survey of in vivo obligate chaperonin-dependent substrates. *EMBO J.* 29, 1552–1564.

(144) Bogumil, D., and Dagan, T. (2010) Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes. *Genome Biol. Evol.* 2, 602–608.

(145) Williams, T. A., and Fares, M. A. (2010) The Effect of Chaperonin Buffering on Protein Evolution. *Genome Biol. Evol.* 2, 609–619.

(146) Warnecke, T., and Hurst, L. D. (2010) GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* 6, 1–11.

(147) Noivirt-Brik, O., Unger, R., and Horovitz, A. (2007) Low folding propensity and high translation efficiency distinguish in vivo substrates of GroEL from other *Escherichia coli* proteins. *Bioinformatics* 23, 3276–3279.

(148) Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25, 117–124.

(149) Takemoto, K., Niwa, T., and Taguchi, H. (2011) Difference in the distribution pattern of substrate enzymes in the metabolic network of *Escherichia coli*, according to chaperonin requirement*. BMC Syst. Biol.* 5, 98.

(150) Newman, M. E. J. (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev.* E 74, 0605087.

(151) Kampinga, H. H., and Craig, E. A. (2010) The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell Biol.* 11, 579–592.

(152) Taipale, M., Jarosz, D. F., and Lindquist, S. (2010) HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nat. Rev. Mol. Cell Biol.* 11, 515–528.

(153) Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49, 169–181.

(154) Gray, M. W., Lukes, J., Archibald, J. M., Keeling, P. J., and Doolittle, W. F. (2010) Cell biology. Irremediable complexity? *Science* 330, 920–921.

(155) Lansbury, P. T. (1999) Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3342–3344.

(156) Bjedov, I., Tenaillon, O., Gérard, B., Souza, V., Denamur, E., Radman, M., Taddei, F., and Matic, I. (2003) Stress-induced mutagenesis in bacteria. *Science* 300, 1404–1409.

(157) Rosenberg, S. M., and Hastings, P. J. (2003) Modulating mutation rates in the wild. *Science* 300, 1382–1383.

(158) Geller, R., Vignuzzi, M., Andino, R., and Frydman, J. (2007) Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes Dev.* 21, 195–205.

(159) Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010) An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* 141, 344–354.

(160) Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., and Barral, Y. (2010) A Role for Codon Order in Translation Dynamics. *Cell* 141, 355–367.

(161) O'Brien, E. P., Vendruscolo, M., and Dobson, C. M. (2012) Prediction of variable translation rate effects on cotranslational protein folding. *Nat. Commun.* 3, 868.

(162) Niwa, T., Kanamori, T., Ueda, T., and Taguchi, H. (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8937–8942.

(163) Horwich, A. L., Apetri, A. C., and Fenton, W. A. (2009) The GroEL/GroES cis cavity as a passive anti-aggregation device. *FEBS Lett.* 583, 2654–2662.

(164) Ferrone, F. A., Hofrichter, J., and Eaton, W. A. (1985) Kinetics of sickle hemoglobin polymerization. II. A double nucleation mechanism. *J. Mol. Biol.* 183, 611–631.

(165) Lorenzen, N., Cohen, S. I. A., Nielsen, S. B., Herling, T. W., Christiansen, G., Dobson, C. M., Knowles, T. P. J., and Otzen, D. (2012) Role of Elongation and Secondary Pathways in S6 Amyloid Fibril Growth. *Biophys. J.* 102, 2167–2175.

5 Publications

5.1 Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes

Bogumil, D., and Dagan, T. (2010) Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes. Genome Biol. Evol. 2, 602–608.

David Bogumil conducted the experiments, performed the statistical analysis and wrote the manuscript.

Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes

David Bogumil, and Tal Dagan*

Institute of Botany III, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

*Corresponding author: E-mail: tal.dagan@uni-duesseldorf.de.

Accepted: 16 July 2010

Abstract

Many proteins require the assistance of molecular chaperones in order to fold efficiently. Chaperones are known to mask the effects of mutations that induce misfolding because they can compensate for the deficiency in spontaneous folding. One of the best studied chaperones is the eubacterial GroEL/GroES system. In *Escherichia coli*, three classes of proteins have been distinguished based on their degree of dependency on GroEL for folding: 1) those that do not require GroEL, 2) those that require GroEL in a temperature-dependent manner, and 3) those that obligately require GroEL for proper folding. The buffering effects of GroEL have so far been observed in experimental regimens, but their effect on genomes during evolution has not been examined. Using 446 sequenced proteobacterial genomes, we have compared the frequency of amino acid replacements among orthologs of 236 proteins corresponding to the three categories of GroEL dependency determined for *E. coli*. Evolutionary rates are significantly correlated with GroEL dependency upon folding with GroEL dependency class accounting for up to 84% of the variation in amino acid substitution rates. Greater GroEL dependency entails increased evolutionary rates with GroEL obligatory proteins (Class III) evolving on average up to 15% faster than GroEL partially dependent proteins (Class II) and 35% faster than GroEL-independent proteins (Class I). Moreover, GroEL dependency class correlations are strictly conserved throughout all proteobacteria surveyed, as is a significant correlation between folding class and codon bias. The results suggest that during evolution, GroEL-dependent folding increases evolutionary rate by buffering the deleterious effects of misfolding-related mutations.

Key words: genome evolution, misfolding, GroEL, codon usage.

Introduction

Chaperones (Ellis 1987), also called heat-shock proteins (HSPs), are essential in both prokaryotes and eukaryotes as they assist protein folding, prevent protein aggregation, and play a crucial role in survival under stress conditions (Young et al. 2004). Moreover, chaperones have been shown to buffer mutational effects both in eukaryotes and in prokaryotes (Rutherford 2003). In *Arabidopsis thaliana*, the reduction of Hsp90 expression level exposes genotype-independent phenotypic variation (Queitsch et al. 2002). In prokaryotes, Hsp60 (GroEL) is essential to organismal fitness under high mutational loads in *Escherichia coli* (Fares et al. 2002; Maisnier-Patin et al. 2005) and in *Buchnera aphidicola* (Moran 1996). Hence in individual organisms, chaperones exert a buffering effect on slightly deleterious mutations, presumably by compensating for decreased folding stability of mutated proteins (Moran 1996; Todd et al. 1996; Fares et al. 2002; Queitsch et al. 2002; Maisnier-Patin et al. 2005; Tokuriki and Tawfik 2009). Is this property widespread in nature and does it affect prokaryote genome evolution?

The chaperone pathway in eubacteria includes a ribosome-bound trigger factor that meets polypeptides as they emerge from the ribosome. The DnaK (Hsp70) and its cochaperone DnaJ may bind alternatively to nascent polypeptides. Subsequently, the GroEL/GroES (Hsp60) chaperonine system operates on a subset of the proteins whose folding requires further energy investment (Young et al. 2004). In *E. coli*, GroEL/GroES is found to interact with about 10% of all soluble proteins (Kerner et al. 2005) and is the only chaperone essential to the bacterium under all tested conditions (Horwich et al. 1993). The GroEL/GroES chaperones are found in all eubacteria except a few highly reduced endosymbionts (Lund et al. 2003). Proteins found in interaction with GroEL in *E. coli* can be classified into three dependency classes (Kerner et al. 2005): GroEL-independent proteins (Class I) fold spontaneously in standard conditions (37 °C)

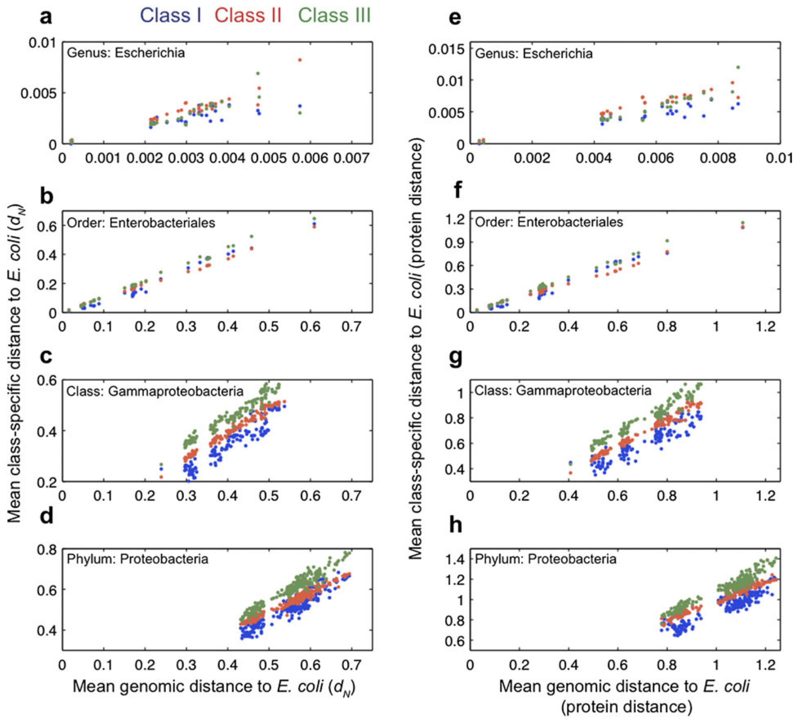


Fig. 1.—Evolutionary rates of proteins in the three GroEL dependency classes within 445 Proteobacteria compared with their *Escherichia coli* strain O157H7 EDL933 ortholog. Each dot in the figure represents the mean distance of all proteins in the same class within the same species from their ortholog in *E. coli* O157H7 EDL933.

and attain on average 55% of their activity independent of chaperones, GroEL, or otherwise. GroEL partially dependent proteins (Class II) require GroEL/GroES assistance, in addition to other chaperons, at 37 °C but do not require GroES at 25 °C, where spontaneous folding is observed. GroEL obligatory proteins (Class III) fail to fold spontaneously at 37 °C and have an obligate requirement for GroEL/GroES in order to attain activity (Kerner et al. 2005). GroEL is known to be a capacitor for slightly deleterious mutations in vitro (Fares et al. 2002; Queitsch et al. 2002; Maisnier-Patin et al. 2005; Tokuriki and Tawfik 2009). If this is also true in nature, Class III proteins should exhibit increased numbers of nonsynonymous substitutions in comparison to Classes I and II.

Materials and Methods

GroEL dependency classes were obtained from Kerner et al. (2005). The Kerner et al. (2005) list contains 249 SWISSPROT accession numbers from various *E. coli* strains. Four proteins that are classified into more than one class were removed. Completely sequenced genomes of 446 Proteobacteria were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/; July 2009 version). Non-proteobacterial taxa were not included in the analysis because we cannot assume that protein interaction with GroEL is conserved in all prokaryotes. In order to use a single reference genome in our analysis, the Kerner et al. (2005) proteins were Blasted (Altschul et al. 1990) on *E. coli* O157H7 EDL933. Proteins that had hits below 98% identical amino acids were curated manually and nine proteins were removed. The remaining proteins distribute as follows: 37 Class I, 120 Class II, and 79 Class III proteins.

Orthologs to *E. coli* strain O157H7 EDL933 proteins in all completely sequenced Proteobacteria were inferred using a reciprocal best Blast hit procedure (Tatusov et al. 1997) with an *e* value $<1 \times 10^{-10}$ cutoff. All orthologous protein pairs were aligned using ClustalW (Thompson et al. 1994). Pairwise alignment reliability was tested using HoT (Landan and Graur 2007), and alignments having column score <90% were excluded. Protein alignments were translated into nucleotide alignments using PAL2NAL (Suyama et al. 2006). Rates of nonsynonymous nucleotide substitutions were calculated by an approximation to maximum likelihood method using yn00 (Yang 2007). Protein distances were calculated by PROTDIST (Felsenstein 2005) using Jones, Taylor, and Thorton (JTT) substitution matrix (Jones et al. 1992). Preferred codons

Table 1
Statistical Tests for Homogeneity of Medians among the GroEL Dependency Classes

Variable	Taxonomic Group	Homogeneity of Medians (*P* value)[a]	Post hoc Comparisons[b]
d_N	Genus: Escherichia	$7.5 \times 10^{-15*}$	I < II, III and II = III[c]
	Order: Enterobacteriales		
	Class: Gammaproteobacteria	$<2.2 \times 10^{-16*}$	I < II < III
	Phylum: Proteobacteria		
Protein distance	Genus: Escherichia	$1.1 \times 10^{-16*}$	I < II, III and II = III
	Order: Enterobacteriales		
	Class: Gammaproteobacteria		
	Phylum: Proteobacteria	$<2.2 \times 10^{-16*}$	I < II < III
CAI	Genus: Escherichia	$<2.2 \times 10^{-16*}$	I > II, III and II = III
	Order: Enterobacteriales		
	Class: Gammaproteobacteria		I > II > III
	Phylum: Proteobacteria		I > II, III and II = III

[a] Using Friedman test.
[b] $\alpha = 0.05$, using Tukey's test.
[c] Roman numbers denote the classes. The notation I < II means that the values of the tested variable are significantly smaller in Class I proteins than in Class II proteins.
**P* value << 0.01.

for each genome and codon adaptation index (CAI) (Sharp and Li 1987) for all genes were calculated using the EMBOSS package (Rice et al. 2000). Amino acid usage and GC content were calculated using an in-house PERL script. Statistical analysis was performed using MatLab statistical toolbox.

To test our hypothesis in different phylogenetic ,we grouped the species in the genome sample into four groups according to their relatedness with *E. coli* strain O157H7 EDL933: 1) Genus: Escherichia, 2) Order: Enterobacterialles, 3) Class: Gammaproteobacteria, and 4) Phylum: Proteobacteria. In order to keep the groups independent, each genome is included in a single group. The genomes are sorted into the groups by their phylogenetic relations with *E. coli*.

Results

To compare nonsynonymous substitution rates among orthologs of the *E. coli* GroEL Class I (37 members), Class II (120 members), and Class III (79 members) proteins, we identified and aligned (Thompson et al. 1994) their orthologs from 446 sequenced proteobacterial genomes. Numbers of nonsynonymous nucleotide substitutions (d_N) (Nei and Gojobori 1986) and amino acid replacements were calculated in pairwise genome comparisons (Yang 2007). For a given genome comparison, the three class-specific mean d_N values were plotted against the mean of all comparisons for the genome pair; this compensates for genome- and lineage-specific differences in substitution rate and nucleotide bias.

Plotting these values at different phylogenetic depths revealed strong and distinct differences in evolutionary rate for the three protein classes, differences which become increasingly apparent with increasing sequence divergence (fig. 1). For intraspecific comparisons within *E. coli* (fig. 1*a*), the differences among the three GroEL dependency classes are not readily visible because of stochastic variation for small d_N values, but they are significant ($P = 7.55 \times 10^{-15}$, using the Friedman test; Zar 1999), with Class I proteins having

Table 2
Explained Variability and Mean Ratios of Class-Specific Values for All Tested Samples

	Genus: Escherichia	Order: Enterobacteriales	Class: Gammaproteobacteria	Phylum: Proteobacteria
d_N				
Explained variability[a]	0.36	0.4	0.87	0.8
Class III/II	0.92	1.06	1.14	1.1
Class III/I	1.1[b]	1.4	1.31	1.18
Protein distance				
Explained variability	0.6	0.3	0.84	0.76
Class III/II	0.87	1.06	1.15	1.1
Class III/I	1.17[b]	1.36	1.35	1.2
CAI				
Explained variability	0.96	0.57	0.48	0.53
Class III/II	0.99	1	0.99	1
Class III/I	0.95	0.98	0.97	0.97

[a] Explained variability was calculated by partial $\eta^2 = \eta^2 = \frac{SS_{\text{treatment}}}{SS_{\text{treatment}} + SS_{\text{error}}}$ with Friedman test.
[b] *Escherichia coli* K12 MG1655 and *E. coli* O157H7 comparisons resulted in zero distance for Class I proteins and were omitted from the calculation.

significantly lower rates than Class II and Class III proteins ($\alpha = 0.05$, using Tukey's post hoc test; Zar 1999). The same test on a larger and ~100-fold more divergent orthologs set from 60 enterics (but excluding *E. coli*) shows a more significant difference in d_N among the GroEL dependency classes ($P < 2.2 \times 10^{-16}$, using Friedman test; fig. 1*b*), with Class I proteins having significantly lower d_N than Class II proteins, and the latter having significantly lower d_N than Class III proteins ($\alpha = 0.05$, using Tukey's post hoc test).

Comparisons within the Gammaproteobacteria (135 genomes; excluding enterics) yielded even more significant correlations (table 1) and furthermore a striking distinction of the three classes (fig. 1*c*). Differences between the GroEL dependency classes account for 87% of the variation between class-specific mean d_N values (table 2). Extending the sample to include 227 Proteobacteria (excluding Gammaproteobacteria) entailed comparisons of greater divergence, with most d_N values exceeding 0.5 substitutions per site (fig. 1*d*), but the significance and the trends remained (table 1), with GroEL dependency class accounting for 80% of the observed differences in class-specific mean d_N (table 2). These correlations held up for GroEL dependency class in amino acid sequence comparisons for the same phylogenetic samples (fig. 1*e*–*h*). At the level of amino acid replacements estimated by JTT (Jones et al. 1992) protein distances for Gammaproteobacteria, Class III proteins evolve on average 15% faster than Class II and 35% faster than Class I proteins (table 2). GroEL folding dependency thus appears to be a major and hitherto undetected determinant of sequence divergence in prokaryotes.

But is the correlation causal? Protein conservation and expression level are known to be correlated (Krylov et al. 2003 ; Drummond et al. 2006; Pál et al. 2006). If chaperon dependency is related to expression level, then it is possible that expression level is the determinant of evolutionary rate differences among the GroEL dependency classes (Warnecke and Hurst 2010). A comparison of protein expression levels measured for *E. coli* strain K12 MG1655 (Lu et al. 2007) shows that these are not equal among the three classes ($P = 2.1 \times 10^{-5}$, using Kruskal–Wallis) with Class I proteins having significantly higher expression levels than Classes II and III proteins, whereas Classes II and III do not differ significantly from each other in their expression levels ($\alpha = 0.05$, using Tukey's post hoc test; fig. 2). To test if protein expression level has any effect on our results, we compared the evolutionary rates among the three GroEL dependency classes while adjusting for the variability in protein expression levels using analysis of covariance (ANCOVA). For the comparison within the genus level and order level, we found significant differences between the three GroEL dependency classes also when protein expression level is considered as the covariate variable (table 3). The ANCOVA was not applicable for the class and phylum levels because the underlying assumptions for that test were not met.

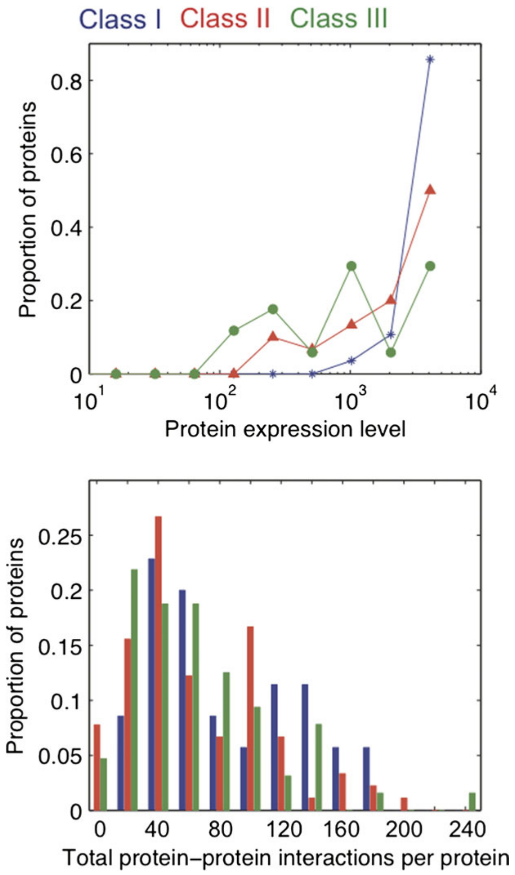


Fig. 2.—Distribution of protein expression levels (Lu et al. 2007) (top) and number of protein-protein interactions (Hu et al. 2009) (bottom) in the three GroEL dependency classes.

Protein expression level has been shown to be positively correlated with the connectivity of a protein within the cellular protein–protein interaction (PPI) network in yeast (von Mering et al. 2002). However, the correlation strength is highly dependent upon the method used to detect interacting proteins (von Mering et al. 2002). Here we tested for difference in PPI frequency among the three dependency classes by using PPI from Hu et al. (2009). We find that the three dependency classes are statistically different in their PPI frequency ($P = 0.049$, using Kruskal–Wallis test) with Class I proteins having a slightly higher frequency of PPIs (median PPI per protein—Class I: 64, Class II: 50; Class III: 52; fig. 2).

We also compared the CAI (Sharp and Li 1987), which is positively correlated, and strongly so, with expression level (Sharp and Li 1987), among orthologs in the three dependency classes at different phylogenetic depths. Class I

Table 3

Statistical Tests for Differences in Evolutionary Rates among the Three GroEL Dependency Classes with a Covariate

Response Variable (y)	Covariate (x)	Taxonomic Group	Pooled Regression[a]	Homogeneity of Slopes among Groups[b]	Homogeneity of Intercepts among Groups[c]
d_N	Protein expression level	Genus: Escherichia	0.026^{*}	0.074	0.0049^{*}
		Order: Enterobacteriales	$6.5 \times 10^{-6**}$	0.52	$<2.2 \times 10^{-16**}$
		Class: Gammaproteobacteria	$<2.2 \times 10^{-16**}$	$<2.2 \times 10^{-16**}$	n.a.
		Phylum: Proteobacteria	$<2.2 \times 10^{-16**}$	$<2.2 \times 10^{-16**}$	n.a.
Protein distance	Protein expression level	Genus: Escherichia	0.0044^{*}	0.15	6.5×10^{-4}
		Order: Enterobacteriales	$1.6 \times 10^{-4**}$	0.49	$<2.2 \times 10^{-16}$
		Class: Gammaproteobacteria	$<2.2 \times 10^{-16**}$	$1.1 \times 10^{-16**}$	n.a.
		Phylum: Proteobacteria	$<2.2 \times 10^{-16**}$	$<2.2 \times 10^{-16**}$	n.a.
d_N	CAI	Genus: Escherichia	$1.3 \times 10^{-9**}$	$5.5 \times 10^{-4**}$	n.a.
		Order: Enterobacteriales	$<2.2 \times 10^{-16**}$	$<2.2 \times 10^{-16**}$	n.a.
		Class: Gammaproteobacteria	$<2.2 \times 10^{-16**}$	$6.1 \times 10^{-6**}$	n.a.
		Phylum: Proteobacteria	$<2.2 \times 10^{-16**}$	0.74	$<2.2 \times 10^{-16**}$
Protein distance	CAI	Genus: Escherichia	$7.7 \times 10^{-13**}$	$<2.2 \times 10^{-16**}$	n.a.
		Order: Enterobacteriales	$<2.2 \times 10^{-16**}$	$5.1 \times 10^{-9**}$	n.a.
		Class: Gammaproteobacteria	$<2.2 \times 10^{-16**}$	$1.9 \times 10^{-13**}$	n.a.
		Phylum: Proteobacteria	$<2.2 \times 10^{-16**}$	0.42	$<2.2 \times 10^{-16**}$

NOTE.—Results of the ANCOVA test and its underlying assumptions (Sokal and Rohlf 1995) are presented. To adjust for overall differences among species, the response variable was divided by the genomic average.

[a] Using F-test for linear relation between the response and covariate $y = ax + b$ testing the null hypothesis H_0: $a = 0$.

[b] Using F-test for equality of slopes among the groups. Each group is fitted with a linear regression $y_{\text{class}} = a_{\text{class}}x_{\text{class}} + b_{\text{class}}$ followed by testing the null hypothesis H_0: $a_{\text{class I}} = a_{\text{class II}} = a_{\text{class III}}$.

[c] Using F-test for equality of intercepts among the groups. This is equivalent to a test for equality of means with the null hypothesis H_0: $\mu_{\text{class I}} = \mu_{\text{class II}} = \mu_{\text{class III}}$.

$*P$ value < 0.05.

$**P$ value $<< 0.01$.

proteins have significantly higher CAI than Classes II and III proteins, whereas CAI values of Class II proteins are either similar (in the order and phylum sets) or slightly increased in comparison to Class III proteins (table 1 and fig. 3). This trend is true not only for *E. coli* (Warnecke and Hurst 2010) but throughout the proteobacteria. Thus, although high expression levels can explain the decreased evolutionary rates for Class I proteins, it cannot explain the increased evolutionary rates in Class III proteins in comparison to Class II proteins. Hence, the difference in evolutionary rates among the three GroEL dependency groups does indeed appear to be attributable to GroEL buffering effects.

Proteins in the three dependency classes are highly dissimilar in their amino acid composition. A comparison of *E. coli* O157H7 EDL933 proteins shows that Class II and Class III proteins comprise significantly more positively charged amino acids (Fujiwara et al. 2010) and less negatively charged amino acids than Class I proteins. No significant difference is found in hydrophobic amino acids or polar uncharged amino acids composition (supplementary table S1, Supplementary Material online). Cysteine and proline usage is significantly higher in Class II and Class III proteins in comparison to Class I proteins. No significant difference in glycine usage among the classes was found (supplementary table S1 and supplementary fig. S1, Supplementary Material online). Genes encoding for Class III proteins are significantly GC richer than Class I proteins (supplementary table S1, Supplementary Material online). This result is attributable to the amino acid usage of Class III proteins, most of them are encoded by GC-rich codons. Repeating this analysis for all orthologs in all phylogenetic depths reveals that the same trends in amino acid usage are general for all tested proteobacteria (supplementary table S2 and supplementary figs. S2–S5, Supplementary Material online). No correlation was found between any of the amino acid usage measures and evolutionary rates (supplementary table S1, Supplementary Material online); hence, the difference in amino acid usage among the GroEL dependency classes may be attributed to the interaction with GroEL (Fujiwara et al. 2010).

Discussion

GroEL can buffer slightly deleterious mutations in experimental setups. In nature this same capacity leads to increased evolutionary rates for GroEL-dependent proteins. It has recently been suggested that protein misfolding has a key role in determining evolutionary rates (Drummond et al. 2005; Drummond and Wilke 2008; Lobkovsky et al. 2010; Warnecke and Hurst 2010). Our results indicate that GroEL-dependent folding is a biological mechanism that can manifest such effects. However, the correlation of GroEL dependency classes with evolutionary rates, protein expression levels, and CAI implies that the promiscuous amino acid substitution regime allowed by the GroEL buffering might not be uniformly distributed within the cellular protein network. The Class I proteins comprise a group of highly

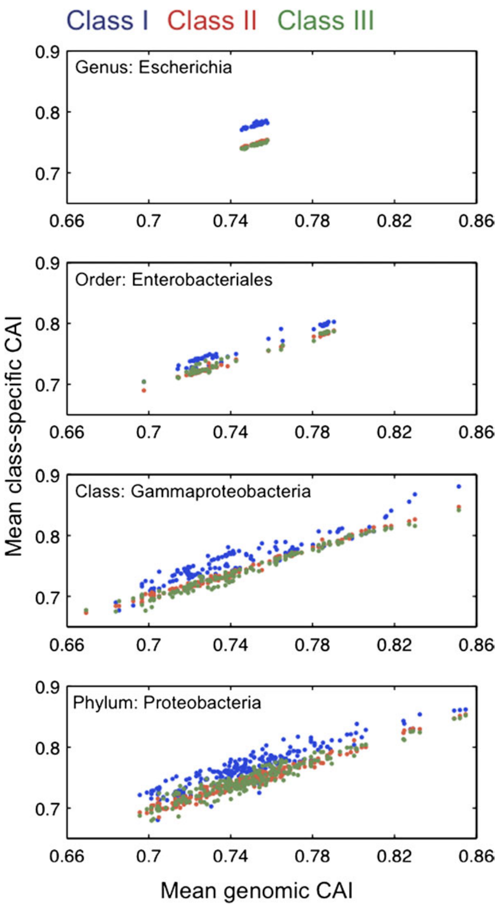


Fig. 3.—CAI of proteins in the three GroEL dependency classes.

conserved, highly expressed proteins having higher CAIs. In contrast, the Class III proteins evolve with an increased evolutionary rate (fig. 1), are expressed at lower levels (fig. 2), and are encoded by less preferred codons (Warnecke and Hurst 2010) (fig. 3). Protein expression level is positively correlated with the number of protein interactions and negatively correlated with dispensability (Pál et al. 2006), whereas CAI is correlated with translation accuracy and efficiency (Drummond and Wilke 2008; Tuller et al. 2010). Hence, proteins that are essential to the cell and that are highly connected in the *E. coli* protein network are not only more conserved but also translated with higher accuracy and tend to fold spontaneously. Conversely, proteins that have a more peripheral role within the cell are more tolerant to increased evolutionary rates and are protected from slightly deleterious mutations by the buffering effect of the GroEL/GroES chaperone.

Supplementary Material

Supplementary figures S1–S6 and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank the German Federal Ministry of Education and Research (BMBF) for financial support. We are thankful to Giddy Landan for his help in statistical analysis and to Bill Martin, Rotem Sorek, and Martin Lercher for their help in refining the manuscript.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 134:341–352.

Ellis RJ. 1987. Proteins as molecular chaperones. Nature. 328:378–379.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. Endosymbiotic bacteria: groEL buffers against deleterious mutations. Nature. 417:398.

Felsenstein J. 2005. PHYLIP (phylogeny inference package). Version 3.6. Seattle (WA): Department of Genome Sciences, University of Washington.

Fujiwara K, Ishihama Y, Nakahigashi K, Soga T, Taguchi H. 2010. A systematic survey of in vivo obligate chaperonin-dependent substrates. EMBO J. 29:1552–1564.

Horwich AL, Low KB, Fenton WA, Hirshfield IN, Furtak K. 1993. Folding in vivo of bacterial cytoplasmic proteins: role of GroEL. Cell. 74:909–917.

Hu P, et al. 2009. Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 7:e96.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation rate matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Kerner MJ, et al. 2005. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. Cell. 122:209–220.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. Proc Natl Acad Sci U S A. 107:2983–2988.

Lu P, Vogel C, Wang R, Yao X, Marcotte EM. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 25:117–124.

Lund PA, Large AT, Kapatai G. 2003. The chaperonins: perspectives from the archaea. Biochem Soc Trans. 31:681–685.

Maisnier-Patin S, et al. 2005. Genomic buffering mitigates the effects of deleterious mutations in bacteria. Nat Genet. 37:1376–1379.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. 93:2873–2878.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Queitsch C, Sangster TA, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. Nature. 417:618–624.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Rutherford SL. 2003. Between genotype and phenotype: protein chaperones and evolvability. Nat Rev Genet. 4:263–274.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Sokal RR, Rohlf FJ. 1995. Biometry. 3rd ed. San Francisco (CA): Freeman.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Todd MJ, Lorimer GH, Thirumalai D. 1996. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. Proc Natl Acad Sci U S A. 93:4030–4035.

Tokuriki N, Tawfik DS. 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature. 459:668–673.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 107:3645–3650.

von Mering C, et al. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. Nature. 417:399–403.

Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. Mol Syst Biol. 6:340.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Young JC, Vishwas RA, Siegers K, Hartl FU. 2004. Pathways of chaperone-mediated protein folding in the cytosol. Nat Rev Mol Cell Biol. 5:781–791.

Zar JH. 1999. Biostatistical analysis. Upper Saddle River (NJ): Prentice Hall.

Associate editor: Takashi Gojobori

5.2 Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate

Bogumil, D., Landan, G., Ilhan, J., and Dagan, T. (2012) Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate. Genome Biol. Evol. 4, 618–625.

David Bogumil conducted the experiments, performed the statistical analysis and wrote the manuscript.

Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate

David Bogumil[1], Giddy Landan[1,2], Judith Ilhan[1], and Tal Dagan[1,]*

[1]Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Germany

[2]Department of Biology & Biochemistry, University of Houston

*Corresponding author: E-mail: tal.dagan@uni-duesseldorf.de.

Accepted: 8 March 2012

Abstract

It has long been known that many proteins require folding via molecular chaperones for their function. Although it has become apparent that folding imposes constraints on protein sequence evolution, the effects exerted by different chaperone classes are so far unknown. We have analyzed data of protein interaction with the chaperones in *Saccharomyces cerevisiae* using network methods. The results reveal a distinct community structure within the network that was hitherto undetectable with standard statistical tools. Sixty-four yeast chaperones comprise ten distinct modules that are defined by interaction specificity for their 2,691 interacting proteins. The classes of interacting proteins that are in turn defined by their dedicated chaperone modules are distinguished by various physiochemical protein properties and are characterized by significantly different protein expression levels, codon usage, and amino acid substitution rates. Correlations between substitution rate, codon bias, and gene expression level that have long been known for yeast are apparent at the level of the chaperone-defined modules. This indicates that correlated expression, conservation, and codon bias levels for yeast genes are attributable to previously unrecognized effects of protein folding. Proteome-wide categories of chaperone–substrate specificity uncover novel hubs of functional constraint in protein evolution that are conserved across 20 fungal genomes.

Key words: codon usage, community structure, networks, protein folding.

Introduction

Chaperones (Ellis 1987), also called heat shock proteins (HSPs), are essential in all living cells as they assist protein folding, prevent protein aggregation, and play a crucial role in survival under stress conditions (Young et al. 2004). Manipulation of chaperone expression has revealed that chaperones have an additional role as capacitors of phenotypic variation (Fares et al. 2002; Queitsch et al. 2002; Rutherford 2003). Inhibition of Hsp90 chaperone function in *Arabidopsis thaliana* exposes genotype-independent phenotypic variation in a similar manner to growth under heat stress conditions (Queitsch et al. 2002). Increasing the expression level of the GroEL (Hsp60) chaperone confers improved fitness in *Escherichia coli* under high mutational loads (Fares et al. 2002). Chaperones can thus buffer the effects of slightly deleterious mutations, presumably by compensating for decreased protein structure stability of mutated proteins (Fares et al. 2002; Queitsch et al. 2002; Rutherford 2003).

Protein interaction with the chaperones for folding impacts the evolvability of substrate proteins (Rutherford 2003; Tokuriki and Tawfik 2009). Overexpression of GroEL/GroES can double the number of accumulating mutations in GroEL substrates in vitro (Tokuriki and Tawfik 2009). Furthermore, the amino acid substitution rate of proteins that depend upon the GroEL for folding in *E. coli* is higher than that of GroEL-independent proteins (Bogumil and Dagan 2010). Here, we study the impact of protein interaction with chaperones on whole-genome evolutionary dynamics. To address this question, we used a network approach to analyze an extensive data set of chaperone–protein interactions assembled by screening for chaperone-associated protein complexes in yeast (Gong et al. 2009). The chaperone repertoire in the *Saccharomyces cerevisiae* proteome consists of 69 molecular chaperones and their co-chaperones, most of which are known to assist the folding or unfolding of proteins in the cell; other chaperones assume diverse cellular functions including translocation across membranes and stabilizing protein–protein interactions (Voos and Röttgers 2003; Young et al. 2004; Kampinga and Craig 2010). The majority of nascent polypeptides in the yeast protein-folding pathway interact with the ribosome-associated complex (RAC) that includes a member of the Hsp70 family and a co-chaperone from the Hsp40 family (J-proteins) (Young et al. 2004; Kampinga and Craig 2010). Some proteins also interact with one or more of the following chaperone classes: prefoldin (PFD), TriC (CCT), and Hsp90 (Young et al. 2004). Most of the proteins encoded in the yeast genome (3,595 of 5,880) interact with at least one chaperone, many of them (2,952) with two or more chaperones (Gong et al. 2009). The present networks uncover hitherto unrecognized modular interactions between chaperone families and their interacting proteins.

Materials and Methods

Data

Data of chaperone interaction repertoire in *S. cerevisiae* were downloaded from Gong et al. (2009). Amino acid usage data, functional assignment, chromosomal location, frequencies of optimal codons, codon adaptation index (CAI), gravy scores (hydropathy index), and aromaticity scores were obtained from the Saccharomyces Genome Database (Cherry et al. 1997). Protein cellular localization was obtained from Huh et al. (2003) and the Gene Ontology database (Ashburner et al. 2000). Secondary structure of all proteins was inferred using PsiPred (Jones 1999). For the calculation of secondary structure usage, a threshold of probability >0.7 was used. Protein expression data were obtained from Ghaemmaghami et al. (2003). For the statistical analysis, the natural log of protein expression was used. Proteins with no expression level information (107) or with zero expression level (1,665) were omitted from the analysis. All statistical analyses were performed using MatLab Statistics toolbox.

Network Modularity Structure

A division of the nodes in the network into modules was obtained by defining a modularity function of each bipartition of the network, as the number of edges within a module minus the expected number of edges in the module. Maximizing this function over all possible divisions using eigenspectrum analysis yields the optimal division of the network into modules (Newman 2006).

Evolutionary Rate

Positional orthology assignments within 20 fungal proteomes were obtained from Wapinski et al. (2007). Open reading frames lacking orthologs (282 in total) were omitted from the analysis. Multiple alignments of all yeast open reading frames with orthologous sequences were reconstructed with MAFFT (Katoh et al. 2005). Phylogenetic trees were reconstructed with PhyML (Guindon and Gascuel 2003) using the best-fit model as inferred by ProtTest 3 (Darriba et al. 2011) using the Akaike information criterion (Akaike 1974) measure. Distances from the *S. cerevisiae* proteins to their orthologs were calculated as the sum of branch lengths. To calculate the relative amino acid substitution rates of substrates, we first *Z*-transformed the distances to the 20 proteomes separately and then averaged the standardized distances over all orthologs.

Results

Modules in the Chaperone–Substrate Interaction Network

In an extensive screening for proteins that interact with each of the 63 chaperones encoded in yeast, Gong et al. (2009) documented a total of 21,687 interactions. The network reconstructed from Gong et al. (2009) data contains 3,595 entities, 3,526 of which are chaperone-interacting proteins (for simplicity termed “substrates” here, yet making no statement about specificity). The remaining 69 entities are chaperones. We designate this as the chaperone-substrate interaction (CSI) network. The network can be fully defined by a matrix, $A = [a_{ij}]69 \times 3{,}595$, with $a_{ij} = 1$ if chaperone *i* and protein *j* interact and $a_{ij} = 0$ otherwise. The chaperones and substrates form two disjoint sets of nodes where interactions between substrate nodes are not allowed because the data reflect the interactions of chaperones with substrate proteins but not other possible interactions among the substrate proteins. The network is thus semi-multipartite, with 9,194 edges of CSIs and 332 edges of chaperone–chaperone interactions (fig. 1). Co-chaperones in our network were found to interact almost exclusively with chaperones.

The CSI network includes five highly connected Hsp70 chaperones that are linked to almost all substrates in the network (Gong et al. 2009). The remaining 64 chaperones interact with fewer proteins, ranging between 2 and 732 substrates per chaperone. Some chaperones interact with a similar set of substrates, thereby forming communities within the network. We examined the community structure in the network by partitioning it into modules using the modularity optimization method (Newman 2006). For each possible bipartition of the network, a modularity function is defined as the observed number of edges within a community minus the expected number. Maximizing this modularity function using its leading eigenvector yields the modules within the network (Newman 2006). Each module is a community of nodes (chaperones and substrates), and each node is assigned to only one community allowing no multiple assignment of a protein to multiple modules.

The result uncovered ten modules that include a total of 64 chaperones and 2,691 substrates, along with 843 lesser (residual) modules that contain a single protein each. The network groups co-chaperones into modules based on their experimental interaction data with the chaperones (Gong et al. 2009). The modules furthermore group together chaperones that interact frequently with common substrates as

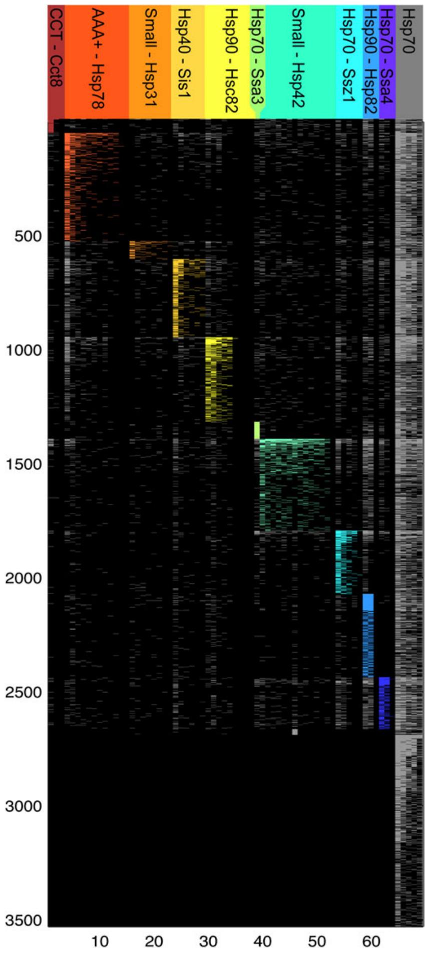


Fig. 1.—The network of CSIs. A graphic representation of the network with chaperones on the *x* axis ($i = 1 \ldots 69$) and substrates on the *y* axis ($j = 1 \ldots 3{,}595$). Cells in the matrix represent a protein–protein interaction between chaperone *i* and substrate *j*. The cells are colored by the module color if both substrate and chaperone are included in the module, and in gray otherwise. Cells of noninteracting proteins are colored in black. Hsp70 group includes the five ungrouped chaperones: Ssb1, Ssa1, Sse1, Ssa2, and Ssb2.

well as those substrates. Five Hsp70 chaperones were not grouped into the ten main modules, forming five single-chaperone modules (Ssa1, Ssa2, Ssb1, Ssb2, and Sse1) (fig. 1). These chaperones are characterized by a promiscuous substrate binding and have many substrates in common (Gong et al. 2009). The remaining 838 singleton modules include proteins that interact solely with the five promiscuous chaperones. We designate the ten main modules by their most connected chaperone. The modules contain between 1 (Hsp70-Ssa3) and 14 (Small-Hsp42) chaperones. The number of substrates folded by each module ranges from 65 (CCT-Cct8) to 485 (AAA+-Hsp78) (supplementary table 1, Supplementary Material online).

The RAC-induced association of Hsp70 family chaperones and J-proteins (Hsp40 family) is clearly evident in the CSI network. For example, the Hsp70-Ssb1 chaperone interacts with 1,044 substrates in total. Of those, 585 (56%) are shared with Hsp40-Ydj1, 483 (46%) with Hsp70-Ssz1, 281 (27%) substrates are shared with Hsp40-Sis1, and 92 (9%) are shared with Hsp40-Zuo1 (Gong et al. 2009). Chaperones Ssb1, Zuo1, and Ssz1 are members of the yeast ribosomal chaperones triad that is anchored to the ribosome and interacts with nascent polypeptides (Gautschi et al. 2001; Conz et al. 2007). No in vivo interactions between Ssb1 and the Hsp40 chaperones Ydj1 or Sis1 have been verified experimentally. Nevertheless, in vitro studies showed that both Ydj1 or Sis1 interact with Ssb1 to determine its specificity for substrate polypeptides (Shorter and Lindquist 2008). The high frequency of common substrates among these chaperones in the Gong et al. (2009) data might indicate that they are associated also in vivo. Three modules (Small-Hsp42, Hsp90-Hsp82, and CCT-Cct8) contain only an Hsp40 chaperone lacking the obligatory partner from Hsp70 family. However, all substrates in these modules also interact with one or more of the five ungrouped promiscuous Hsp70 chaperones. Two modules, Hsp70-Ssa3 and Hsp70-Ssa4, include only an Hsp70 chaperone lacking an Hsp40 partner. Substrates in those two modules interact with various Hsp40 chaperones and with the Ydj1, which has no substrate specificity (Kampinga and Craig 2010), as the most common interactor. Two modules include members of both TriC and PFD chaperone families, whereas three modules include only a TriC chaperone and one module only a PFD chaperone (supplementary table 1, Supplementary Material online).

Members within the modules are not restricted to a certain cellular localization (supplementary fig. 1, Supplementary Material online). This result conforms with the high abundance of interactions between chaperones and substrates that are localized in different cell compartments as reported in various protein–protein interaction databases (70% in Gong et al. (2009) data used here, 66% in BioGrid [ver. 3.1.77], Stark et al. 2006, and 67% in Strings [ver. 8.3], Szklarczyk et al. 2011). This indicates that protein folding and function do not always occur in the same compartment. Module Hsp90-Hsc82 is, however, enriched with chaperones localized in the mitochondrion (5 of 9; supplementary table 1, Supplementary Material online). The module includes Hsp60 and Hsp10 that interact to fold proteins in the mitochondrion (Rospert et al. 1993). These two chaperones are homologous to the eubacterial GroEL/GroES chaperonin system (Gupta 1995). Furthermore, the Hsp70 (Ssc1) and Hsp40 (Mdj2) chaperones in this module are known to be localized in the mitochondrion (supplementary

Table 1
Comparison of Substrate Properties among the Modules

	Variable	As Is[a]	Random	Correlation with Expression Level in the Network[b]
Expression	Expression level	2.22×10^{-16}**	0.62	—
	CAI	2.38×10^{-06}**	0.37	0.54**
	Optimal codons	1.18×10^{-05}**	0.76	0.53**
Secondary structure	Alpha helix	0.0067**	0.08	0.02
	Coiled coils	0.0256**	0.4	0.21**
	Beta sheets	0.0833	0.53	0.21**
Physiochemical properties	Protein length	4.13×10^{-09}**	0.94	−0.17**
	Hydrophobic amino acids	0.2177	0.23	0.18**
	Negative amino acids	0.0008**	0.56	0.08**
	Positive amino acids	0.5682	0.72	−0.06**
	Polar amino acids	0.0081**	0.83	−0.31**
	Aromaticity index	0.0017**	0.43	−0.04**
	Gravy	0.171	0.58	0.14**
Amino acid frequencies	Alanine	6.60×10^{-07}**	0.89	0.36**
	Arginine	0.3581	0.8	−0.09**
	Asparagine	0.0384*	0.58	−0.27**
	Aspartate	4.71×10^{-05}**	0.08	0.03
	Cysteine	0.5354	0.23	−0.09**
	Glutamine	0.0064**	0.87	−0.08**
	Glutamate	0.2669	0.97	0.09**
	Glycine	0.0172**	0.24	0.25**
	Histidine	0.4528	0.07	−0.10**
	Isoleucine	0.0027**	0.11	−0.06**
	Leucine	0.0031**	0.47	−0.08**
	Lysine	0.4807	0.75	0.03**
	Methionine	0.3369	0.61	−0.08**
	Phenyl-alanine	0.0012**	0.48	−0.04**
	Proline	0.0074**	0.43	−0.07**
	Serine	0.0417*	0.07	−0.29**
	Threonine	0.4651	0.72	−0.05**
	Tryptophan	0.0612	0.48	0.03
	Tyrosine	0.0586	0.31	0.02
	Valine	0.0185**	0.27	0.27**
Evolutionary rate	Substitution rate	2.15×10^{-06}**	0.36	−0.42**
	% Identical amino acids	1.35×10^{-07}**	0.81	0.47**
	Substitutions per site	2.58×10^{-07}**	0.75	−0.46**

[a] Using Kruskal–Wallis test for equality of median ranks with the null hypothesis, H_0: $\mu_{module1} = \mu_{module2} = \ldots = \mu_{module10}$.
[b] Using Spearman rank correlation coefficient.
[*] P value < 0.05.
[**] P value < 0.05 using false discovery rate test for multiple comparisons.

table 1, Supplementary Material online) (Huh et al. 2003). Notably, the Hsp90-Hsc82 module is lacking both PFD and TriC chaperones, which are homologous to archaeal chaperones (Hartl and Hayer-Hartl 2009). The chaperone repertoire of this module suggests that it is of mitochondrial origin, reflecting a functional eubacterial unit within the yeast proteome (Esser et al. 2004).

Module Expression and Biochemical Properties

Substrate expression level as measured by protein molecules per cell (Ghaemmaghami et al. 2003) is significantly different among the ten modules (table 1). Substrates in modules Hsp70-Ssa4, Hsp90-Hsp2, and Hsp70-Ssz1 are expressed in the lowest level. Substrates in modules AAA+-Hsp78 and CCT-Cct8 are highly abundant in the cell (fig. 2). Substrates that interact only with the promiscuous Hsp70 chaperones have a higher expression level than substrates within the modules ($P = 1.35 \times 10^{-58}$, using one-sided Kolmogorov–Smirnov). Yeast proteins that are missing from the CSI network have a significantly lower expression level than connected proteins ($P = 2.8 \times 10^{-62}$, using one-sided Kolmogorov–Smirnov). This suggests that those proteins might interact with chaperones but were so far not detected in surveys for chaperone interactors, possibly due to their low expression level. Chaperone expression level shows no

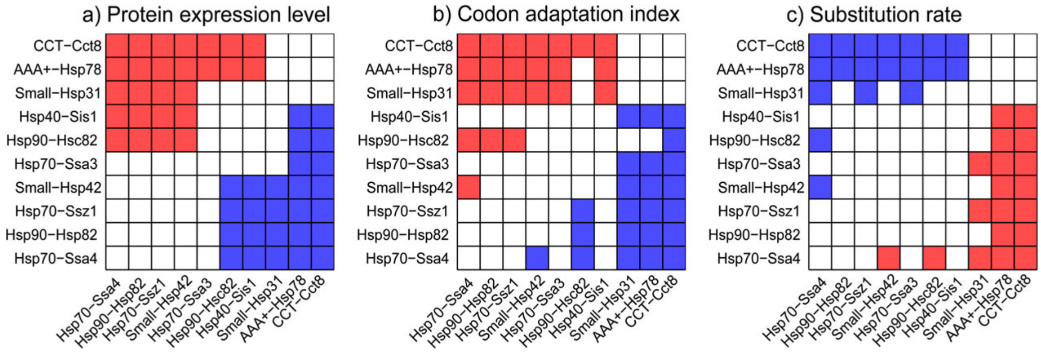


Fig. 2.—Comparison of expression level (*a*), codon adaptation index (*b*), and relative amino acid substitution rates (*c*) among the modules. A matrix representation of post hoc multiple comparison results ($\alpha = 0.05$, using Tukey test). Cell a_{ij} in the matrix is colored red if the corresponding variable module $i >$ module j, blue if module $i <$ module j, and white if no significant difference between the modules was found.

significant differences across the ten modules ($P = 0.051$, using Kruskal–Wallis).

Protein expression and encoding by preferred codons are known to be positively correlated (Sharp and Li 1987). This correlation is apparent also in the CSI network, where substrate expression level is positively correlated with CAI (table 1). A comparison of codon usage among the modules—measured by the CAI (Sharp and Li 1987)—reveals significant difference across the modules (table 1), with modules Hsp70-Ssa4, Hsp90-Hsp82, and Hsp70-Ssz1 having the lowest CAI values and modules AAA+-Hsp78 and CCT-Cct8 having the highest CAIs (fig. 2). A randomization of protein module classification eliminates the significant CAI differences across the modules (table 1). A pairwise comparison of substrate expression level and CAI between the modules reveals that the correlation between these two properties is apparent at the modules level with highly expressed modules having high CAI values and vice versa (fig. 2).

Substrates in the ten modules vary substantially in their physiochemical properties. The secondary structure of substrates—measured by the proportion of alpha helixes and coiled coils—differs significantly among the modules (table 1). Substrates in module Hsp70-Ssz1 are enriched with coiled coil, whereas substrates in module Small-Hsp31 are enriched with alpha helixes (supplementary fig. 1, Supplementary Material online). No significant difference in the proportion of beta-sheet structures was found among the modules (table 1). The amino acid usage of most hydrophobic amino acids differs significantly between the modules (including Ala, Ile, Leu, Phe, and Val) as well as the usage of the negatively charged amino acid Asp (table 1; supplementary fig. 2, Supplementary Material online). Of the polar amino acids, only Gln usage is significantly different across the modules, with substrates in module Hsp70-Ssz1 encoding the highest Gln content. Phe is the only aromatic amino acid whose content varies across the modules (table 1). Substrates in the modules are significantly different in their aromaticity index with substrates in Small-Hsp31 encoding the lowest content of aromatic amino acids (table 1; supplementary fig. 2, Supplementary Material online). Substrate protein length is significantly different among the modules (table 1). The shortest substrates are found in modules AAA+-Hsp78, Small-Hsp31, and Hsp70-Ssa3 and the longest substrates in module Small-Hsp42 (supplementary fig. 2, Supplementary Material online). Randomizing the module classification of substrates eliminated the significant differences among the modules for all of the substrate properties mentioned above (table 1). Furthermore, none of these protein biochemical properties is correlated with protein expression level within the network (table 1).

No clear enrichment for substrate functional category, cellular localization, chromosomal location (supplementary fig. 1, Supplementary Material online), protein domain (supplementary table 2, Supplementary Material online), or sequence motif (supplementary table 3, Supplementary Material online) was found among the modules.

Module Evolutionary Dynamics

To test the impact of protein interaction with the chaperones on protein evolution, we compared substrate amino acid substitution rate among the modules. Phylogenetic trees were reconstructed from a multiple sequence alignment of *S. cerevisiae* substrate proteins with their positional ortholog from among 20 sequenced fungal genomes (Wapinski et al. 2007). A comparison of relative amino acid substitution rates among substrates in the ten modules revealed significant differences across the modules (table 1). Randomizing the module classification of substrates eliminates the differences in evolutionary rate among the modules (table 1). Ranking the modules from slow to fast by their relative substrate amino acid substitution rates shows that modules AAA+-Hsp78, CCT-Cct8, and Small-Hsp31 evolve with the

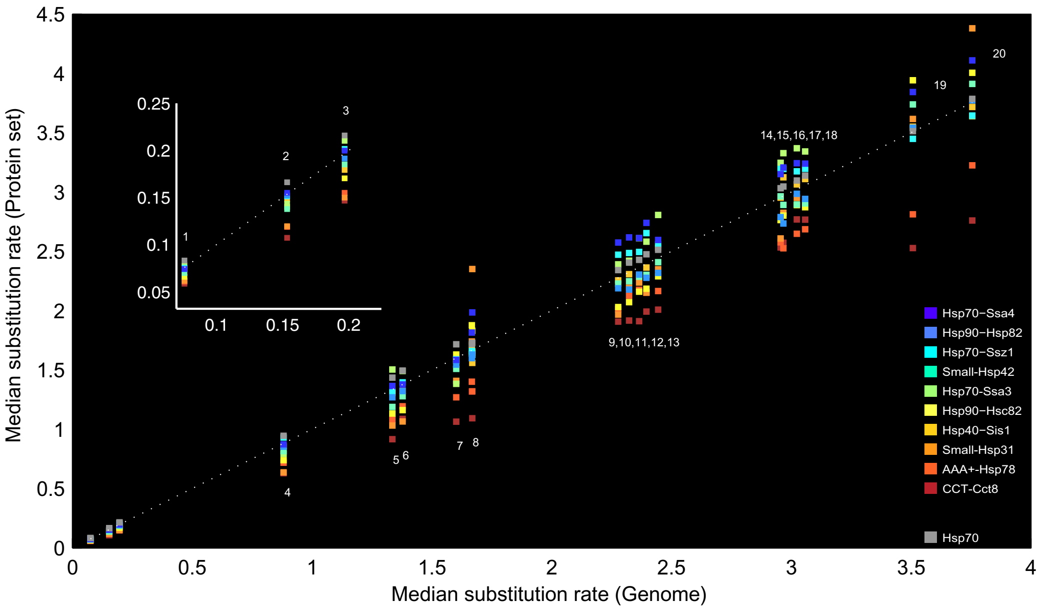


Fig. 3.—Evolutionary distances of yeast substrates in the ten modules compared with their positional ortholog in 20 fungal species. The *x* axis shows the variation of amino acid substitution rates within different fungal genomes in comparison with yeast. The *y* axis shows the rate variation among proteins in the different modules within the same genome. Module colors correspond to the ranking by substrate expression levels with highly expressed modules in red shades and lowly expressed modules in blue shades. Hsp70 group includes five ungrouped chaperones: Ssb1, Ssa1, Sse1, Ssa2, and Ssb2. Arabic numerals correspond to fungal species: 1) *Saccharomyces paradoxus*, 2) *Saccharomyces mikatae*, 3) *Saccharomyces bayanus*, 4) *Saccharomyces castellii*, 5) *Kluyveromyces Lactis*, 6) *Ashbya gossypii*, 7) *Kluyveromyces waltii*, 8) *Lachancea kluyveri*, 9) *Candida glabrata*, 10) *Candida guilliermondii*, 11) *Candida albicans*, 12) *Candida tropicalis*, 13) *Lodderomyces elongosporus*, 14) *Yarrowia lipolytica*, 15) *Aspergillus nidulans*, 16) *Neurospora crassa*, 17) *Schizosaccharomyces pompe*, 18) *Schizosaccharomyces japonicus*, 19) *Debaryomyces hansenii*, and 20) *Candida parapsilosis*.

slowest rates, whereas modules Hsp40-Sis1, Hsp70-Ssa3, and Hsp70-Ssa4 evolve with the highest rates. Substrates in the fastest module (Hsp70-Ssa3) evolve on average 15.6% faster than substrates in the slowest module (CCT-Cct8). Chaperones in the ten modules evolve in similar evolutionary rates ($P = 0.12$, using Kruskal–Wallis).

A comparison of module ranking at the species level reveals that module ranking is conserved during evolution (fig. 3). Substrates in the slowest and fastest modules maintain a similar ranking in almost all compared genomes. The conservation of intermediate module ranking varies to a larger extent. Module ranking is mostly diverged in species that are distantly related to yeast such as *Debaryomyces hansenii* and *Candida parapsilosis*. The intra-*Saccharomyces* comparison shows that substrates interacting exclusively with the five ungrouped Hsp70 chaperones evolve at the fastest rates; in more distantly related fungi, these proteins evolve at rates that are comparable to the fastest modules. Species where the module ranking is conserved (e.g., *S. paradoxus* and *S. mikatae*) are expected to have a CSI network that is similar to that of yeast (fig. 3).

Amino acid substitution rate and protein expression level are known to be inversely correlated at the genome level (Grantham et al. 1981; Pál et al. 2001, 2006; Krylov et al. 2003; Drummond et al. 2005). This correlation is observed also in the CSI network, where substrate expression level is negatively correlated with evolutionary rate (table 1). A comparison between module ranking by evolutionary rate with that of expression level shows that modules that are highly expressed are also the modules that evolve with the slowest substitution rates. Conversely, substrates in modules have the lowest expression levels and evolve in the highest substitution rates (fig. 2). A comparison of the relative amino acid substitution rates among the ten modules while adjusting for the variability in protein expression level reveals that the effect of expression level could not be rejected ($P = 0.56$, using analysis of covariance; $P_{\text{linearity}} = 2.48 \times 10^{-104}$; $P_{\text{slopes homogeneity}} = 0.27$).

Discussion

Chaperones are major hubs within the eukaryotic protein–protein interaction network (Gong et al. 2009). The multiplicity of interacting partners imposes a strong functional constraint on the evolution of hub proteins (Fraser et al. 2002). Moreover, multiple substrates of a certain chaperone evolve under the constraint to interact with that single chaperone. This can explain the similarity in biochemical properties and secondary structure elements among proteins that interact with common chaperones. The differences in substrate physiochemical properties across the modules are probably due to the different structures required for the interaction with the different chaperones.

Notably, the two Hsp70 paralogs Ssb1 and Ssb2 that differ in only two adjacent amino acids (C434V and A435S) were not grouped into the same module, rather each has its independent module. Interaction data of Gong et al. (2009) reveal that they have a different substrate repertoire. Ssb1 interacts with 2,756 (49%) of the substrates in our network; Ssb2 is associated with 1,064 (19%) substrates, and 899 (87%) of them are common with Ssb1 (Gong et al. 2009). The difference in the interaction regime of these two paralogs may be due to the difference in their expression level. Under standard conditions (Ghaemmaghami et al. 2003), Ssb1 is expressed in 170,000 copies in the cell, and Ssb2 is expressed in 104,000 copies. Hence by chance alone, it is more likely that potential Hsp70 substrates will interact more frequently with Ssb1 rather than Ssb2. Substrate specificity in Ssb2 interactions, if exists, is probably determined by chaperone and substrate coexpression or by their specificity to multiprotein complexes (e.g., the RAC complex).

Our analysis reveals that highly and lowly expressed proteins interact with different chaperones. Protein amino acid composition and secondary structure are known to impact the rate of protein folding and structural stability (Dobson 2003; Yang et al. 2010). Protein interaction with the chaperones lowers the energetic barrier for protein folding into the functional conformation (Hartl and Hayer-Hartl 2009). Thus, the evolution of protein–chaperone interaction is expected to depend upon the protein propensity to fold spontaneously. Chaperone-mediated folding ensures proper functional conformation, but it costs both time and energy. For example, protein folding by the GroEL/GroES chaperonin system in *E. coli* takes about 10 s and consumes seven adenosine triphosphate molecules (Horwich et al. 2009). It is therefore probably advantageous to have a subset of proteins that are less dependent upon chaperones for folding. If energetic efficiency is a selective constraint, this subset is likely to be defined by high expression levels and short response time. The spectrum of chaperone interaction with protein substrates can vary. For example, the GroEL/GroES chaperonin system in *E. coli* interacts with both casual and obligatory substrates. Casual interactors bind to GroEL in vivo but can also gain functional activity independent of GroEL in vitro (Kerner et al. 2005). Casual GroEL substrates have significantly higher expression level than obligatory substrates (Bogumil and Dagan 2010), consistent with the results presented here, which suggest that protein abundance within the cell largely determines the kind and mode of interaction with the chaperones for folding.

Protein expression level is known to be positively correlated with the usage of preferred codons (Sharp and Li 1987) and negatively correlated with evolutionary rate (Grantham et al. 1981; Pál et al. 2001, 2006; Krylov et al. 2003; Drummond et al. 2005). Current theories to explain these correlations evoke either poorly specified network properties of proteins (Fraser et al. 2002) or the specific effects of amino acid misincorporation during protein translation (Drummond et al. 2005; Drummond and Wilke 2008; Warnecke and Hurst 2010). Our results show that dividing the yeast proteins into modules by their chaperone interactions also captures the above correlations. The ten modules are significantly different in terms of each of these three properties, yet the 3-fold correlation prevents naming any one of the three measures as the leading causal effect of substrate–chaperon interactions. The question that remains is how protein interaction with the chaperones is related to protein expression level and codon adaptation. Considering the function of yeast chaperones, the majority of interactions in the CSI network correspond to chaperone-mediated protein folding. We suggest that the correlation between expression level and codon usage stems from the requirement for synchronization between protein translation and folding. Recently, it was shown that codon usage distribution along the protein sequence plays a role in protein translation speed (Cannarozzi et al. 2010; Tuller et al. 2010). Proteins that require chaperones have to be translated at a speed that fits the time required for chaperone recruitment (i.e., chaperone abundance and turnover rate), otherwise the protein will fold spontaneously into the wrong conformation, thereby forming aggregates that hinder the cell viability (Geiler-Samerotte et al. 2011). Proteins that can fold spontaneously into their functional conformation are free from that constraint and can be translated at a higher speed. However, with increasing translation speed, accuracy becomes more important, so that proteins that are translated at high speed should be more conserved (Drummond and Wilke 2008). The involvement of chaperones and folding in the yeast correlations between rates, codon bias, and expression introduces new perspectives on the issue.

Supplementary Material

Supplementary figures 1 and 2 and tables 1–3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the German Science Foundation (Deutsche Forschungsgemeinschaft) and a European Research Council grant NETWORKORIGINS to W. Martin. G.L. was supported by US National Library of Medicine grant LM010009-01. Computational support and infrastructure were provided by the Zentrum für Informations- und

Medientechnologie (ZIM) at the Heinrich-Heine-University of Duesseldorf. We are thankful to W. Martin, M. J. Lercher, and E. Shakhnovich for their help in refining the manuscript.

Literature Cited

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automat Contr. 19:716–723.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Bogumil D, Dagan T. 2010. Chaperonin-dependent accelerated substitution rates in prokaryotes. Genome Biol Evol. 2:602–608.

Cannarozzi G, et al. 2010. A role for codon order in translation dynamics. Cell 141:355–367.

Cherry JM, et al. 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 387:67–73.

Conz C, et al. 2007. Functional characterization of the atypical Hsp70 subunit of yeast ribosome-associated complex. J Biol Chem. 282:33977–33984.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1175.

Dobson CM. 2003. Protein folding and misfolding. Nature 426:884–890.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Ellis RJ. 1987. Proteins as molecular chaperones. Nature 328:378–379.

Esser C, et al. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21:1643–1660.

Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E. 2002. GroEL buffers against deleterious mutations. Nature 417:398.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.

Gautschi M, et al. 2001. RAC, a stable ribosome-associated complex in yeast formed by the DnaK-DnaJ homologs Ssz1p and zuotin. Proc Natl Acad Sci U S A. 98:3762–3767.

Geiler-Samerotte KA, et al. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A. 108:680–685.

Ghaemmaghami S, et al. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.

Gong Y, et al. 2009. An atlas of chaperone-protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. Mol Syst Biol. 5:275.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9:43–74.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Gupta RS. 1995. Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol Microbiol. 15:1–11.

Hartl FU, Hayer-Hartl M. 2009. Converging concepts of protein folding in vitro and in vivo. Nat Struct Mol Biol. 16:574–581.

Horwich AL, Apetri AC, Fenton WA. 2009. The GroEL/GroES cis cavity as a passive anti-aggregation device. FEBS Lett. 583:2654–2662.

Huh WK, et al. 2003. Global analysis of protein localization in budding yeast. Nature 425:686–691.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195–202.

Kampinga HH, Craig EA. 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. Nat Rev Mol Cell Biol. 11:579–592.

Katoh K, Kuma KI, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kerner MJ, et al. 2005. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. Cell 122:209–220.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.

Newman MEJ. 2006. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E. 74:036104.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Queitsch C, Sangster TA, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. Nature 417:618–623.

Rospert S, et al. 1993. Identification and functional analysis of chaperonin 10, the groES homolog from yeast mitochondria. Proc Natl Acad Sci U S A. 90:10967–10971.

Rutherford SL. 2003. Between genotype and phenotype: protein chaperones and evolvability. Nat Rev Genet. 4:264–274.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Shorter J, Lindquist S. 2008. Hsp104, Hsp70 and Hsp40 interplay regulates formation, growth and elimination of Sup35 prions. EMBO J. 27:2712–2724.

Stark C, et al. 2006. Biogrid: a general repository for interaction datasets. Nucleic Acids Res. 34:535–539.

Szklarczyk D, et al. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39:561–568.

Tokuriki N, Tawfik DS. 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459:668–673.

Tuller T, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354.

Voos W, Röttgers K. 2003. Molecular chaperones as essential mediators of mitochondrial biogenesis. Biochim Biophys Acta. 1592:51–62.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61.

Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. Mol Syst Biol. 6:340.

Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol. 6:421.

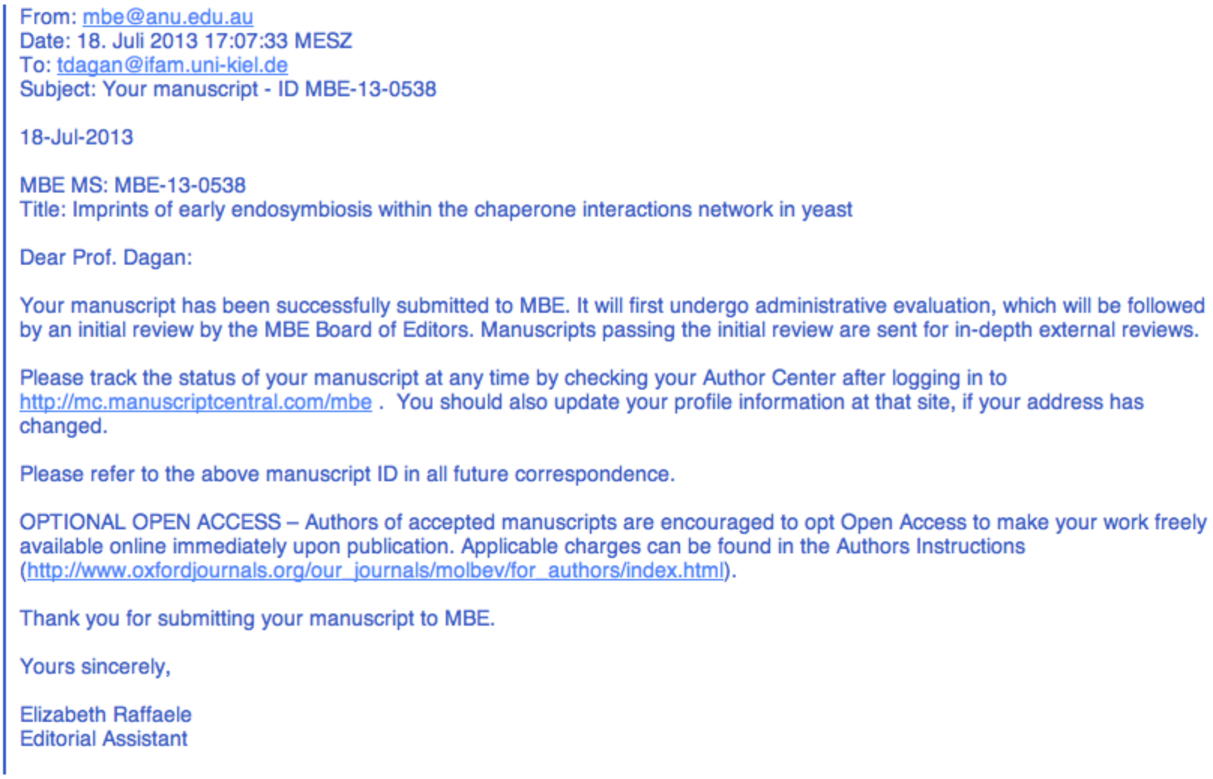
Young JC, Agashe VR, Siegers K, Hartl FU. 2004. Pathways of chaperone-mediated protein folding in the cytosol. Nat Rev Mol Cell Biol. 5:781–791.

Associate editor: Eugene Koonin

5.3 Imprints of early endosymbiosis within the chaperone interactions network in yeast

Bogumil, D., Alvarez-Ponce, D., Landan, G., McInerney, J.O., and Dagan, T. (2013) Imprints of early endosymbiosis within the chaperone interactions network in yeast. *Submitted.*

David Bogumil conducted the experiments, performed the statistical analysis and wrote the manuscript.



From: mbe@anu.edu.au
Date: 18. Juli 2013 17:07:33 MESZ
To: tdagan@ifam.uni-kiel.de
Subject: Your manuscript - ID MBE-13-0538

18-Jul-2013

MBE MS: MBE-13-0538
Title: Imprints of early endosymbiosis within the chaperone interactions network in yeast

Dear Prof. Dagan:

Your manuscript has been successfully submitted to MBE. It will first undergo administrative evaluation, which will be followed by an initial review by the MBE Board of Editors. Manuscripts passing the initial review are sent for in-depth external reviews.

Please track the status of your manuscript at any time by checking your Author Center after logging in to http://mc.manuscriptcentral.com/mbe . You should also update your profile information at that site, if your address has changed.

Please refer to the above manuscript ID in all future correspondence.

OPTIONAL OPEN ACCESS – Authors of accepted manuscripts are encouraged to opt Open Access to make your work freely available online immediately upon publication. Applicable charges can be found in the Authors Instructions (http://www.oxfordjournals.org/our_journals/molbev/for_authors/index.html).

Thank you for submitting your manuscript to MBE.

Yours sincerely,

Elizabeth Raffaele
Editorial Assistant

Imprints of early endosymbiosis within the chaperone interactions network in yeast

David Bogumil[1], David Alvarez-Ponce[2,3], Giddy Landan[1], James O. McInerney[2,4], Tal Dagan[1]

[1] Institute of Microbiology, Christian-Albrechts-University of Kiel, Kiel, Germany

[2] Department of Biology, National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

[3]Current address: Integrative and Systems Biology Laboratory, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Valencia, Spain.

[4]Current address: Center for Communicable Disease Dynamics, Harvard School of Public Health, 677 Huntington Avenue, Boston 02115, Massachusetts, USA.

Abstract

Eukaryotic genomes are mosaics of genes acquired from their prokaryotic ancestors, the eubacterial endosymbiont that gave rise to the mitochondrion and its archaebacterial host. Genomic footprints of the prokaryotic merger at the origin of eukaryotes are still discernable in eukaryotic genomes, where gene expression and function correlate with their prokaryotic ancestry. Molecular chaperones are essential in all domains of life as they assist the functional folding of their substrate proteins and protect the cell against the cytotoxic effects of protein misfolding. Eubacteria and archaebacteria code for slightly different chaperones, comprising distinct protein folding pathways. Here we study the evolution of the eukaryotic protein folding pathways following the endosymbiosis event. A phylogenetic analysis of all 69 chaperones encoded in the *Saccharomyces cerevisiae* genome revealed 26 chaperones of eubacterial ancestry, 11 of archaebacterial ancestry, 10 of ambiguous prokaryotic ancestry and 22 that may represent eukaryotic innovations. Several chaperone families (e.g., Hsp90 and Prefoldin) trace their ancestry to only one prokaryote group, while others, such as Hsp40 and Hsp70, are of mixed ancestry, with members contributed from both prokaryotic ancestors. Analysis of the yeast chaperone–substrate interaction network revealed no preference for interaction between chaperones and substrates of the same origin. Our results suggest that the archaebacterial and eubacterial protein folding pathways have been reorganized and integrated into the present eukaryotic pathway. The highly integrated chaperone system of yeast is a manifestation of the central role of chaperone-mediated folding in maintaining cellular fitness. Most likely, both archaebacterial and eubacterial chaperone systems were essential at the very early stages of eukaryogenesis, and the retention of both may have offered new opportunities for expanding the scope of chaperone-mediated folding.

Introduction

The symbiogenic model for the origin of eukaryotes posits that eukaryotes arose via a symbiotic association of two distantly related prokaryotes (Sagan 1967; Rivera and Lake 2004; Embley and Martin 2006; Pisani et al. 2007; Lane 2009; Alvarez-Ponce et al. 2013). Opinions about the precise taxonomic classification and metabolic capacities of the prokaryote involved are still divided, however there is a wide agreement among scientists that the host was an archaebacterium (Martin and Müller 1998; Cox et al. 2008; Williams et al. 2012) and the endosymbiont was an alpha-proteobacterium (Gray et al. 1999; Gabaldón and Huynen 2003; Esser et al. 2004). The eubacterial endosymbiont subsequently evolved into the mitochondrion organelle, a process that was accompanied by a massive DNA transfer from the symbiont into the host genome, the evolution of a mitochondrial protein import apparatus, a drastic miniaturization of the mitochondrial genome, and an increased complexity of the nuclear genome (Martin and Herrmann 1998; Martin 2003; Timmis et al. 2004). Phylogenomic studies show, accordingly, that eukaryotic genomes are a mosaic of genes of eubacterial and archaebacterial ancestry (Esser et al. 2004; Pisani et al. 2007; Thiergart et al. 2012; Alvarez-Ponce et al. 2013).

Evolutionary analysis of genes in the model eukaryote *Saccharomyces cerevisiae* reveals that about 37% of the genes can be traced back to either an archaebacterial or a eubacterial ancestor (Cotton and McInerney 2010). Thus, eukaryotic innovations probably account for a sizeable fraction of eukaryotic genomes. Yet, the proportion of eukaryotic genes of demonstrable prokaryotic origin is quite substantial considering the complications involved in this kind of analysis. The long divergence time elapsed since the symbiotic event limits our ability to detect prokaryotic homologs to some prokaryote-derived proteins and reduces the accuracy of phylogenetic inference for others. Furthermore, lateral gene transfer events between the eubacterial and archaebacterial lineages (e.g., Deppenmeier et al. 2002; Large and Lund 2009; Williams et al. 2010; Nelson-Sathi et al. 2012) may have obscured the genetic record of the symbiosis event, leading to an ambiguous classification of eukaryotic genes.

The chimerical origin of eukaryotic genomes is imprinted in the functional role of proteins within the cell. Many proteins that perform an informational function (e.g., replication, transcription and translation) are of archaebacterial origin while many genes of eubacterial origin perform operational functions (e.g., metabolism, amino acid synthesis, and regulatory genes) (Rivera et al. 1998; Esser et al. 2004; Cox et al. 2008; Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011; Alvarez-Ponce et al. 2013). Eukaryotic genes of archaebacterial origin are more essential regardless of the bias towards informational functions (Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011). Furthermore, the eukaryotic protein-protein interaction network still bears the markings of a chimerical ancestry, with proteins from the same origin – archaebacterial or eubacterial – being interconnected at a frequency that is significantly above the expected by chance (Alvarez-Ponce and McInerney 2011). Thus, when considered as a whole, the eukaryotic proteome can be described as a partially integrated version of two ancestral ingredients.

In this study we have set forth to examine the evolution of the eukaryotic protein folding pathway in light of the symbiogenic model. Molecular chaperones are proteins that assist the folding and unfolding of other proteins, as well as the complex assembly and stabilization of protein and nucleic acids interactions (Hartl and Hayer-Hartl 2009; Large et al. 2009). Chaperones often function in assembly-line-like pathways where various chaperones interact consecutively with the same substrate driving the transition of the newly synthetized peptide into a functional protein (Young et al. 2004). Chaperones are essential in all living organisms and have been shown to play a role as capacitors of phenotypic variation (Rutherford and Lindquist 1998; Queitsch et al. 2002) and drivers of increased fitness within organisms facing a high mutational load (Fares et al. 2002; Maisnier-Patin et al. 2005). Furthermore, their function as biochemical mediators of protein assembly played an important role in shaping genomic landscapes (Bogumil and Dagan 2010; Williams and Fares 2010; Bogumil et al. 2012). The utility of molecular chaperones is thought to be constrained by a delicate balance between their help in mitigating the effects of protein misfolding and the slower rate of protein production and maturation of their substrate (Bogumil and Dagan 2012). Archaebacteria and eubacteria harbor slightly different repertoires of chaperone families. The Hsp40 and Hsp70 chaperone families are present in both domains (Macario et al. 1991; Macario et al. 1993) whereas other chaperone systems, such as chaperonins, differ in their composition and assembly.

Here we study the extent to which the chimeric origin of eukaryotes is still detectable in the eukaryotic protein folding pathway of contemporary genomes. We infer the ancestry of yeast chaperones and their substrates, examine the yeast chaperone repertoire and use a network approach to study the relationship between chaperones and their substrates in light of their origin.

Results

Prokaryotic ancestry of *S. cerevisiae* proteins

To determine the prokaryotic origin of yeast proteins we searched for their prokaryotic homologs among 82 archaebacterial and 1,074 eubacterial genomes. A total of 1,230 yeast proteins had detectable homologs in one or more prokaryotic genomes. The remaining proteins did not manifest detectable homology with prokaryotic proteins and we therefore consider them to be eukaryotic innovations. A total of 686 phylogenetic trees were reconstructed for yeast proteins having more than three homologs belonging to both archaebacteria and eubacteria. Yeast proteins were classified according to the prokaryotic domain within which they branch. Our analysis revealed 289 proteins of archaebacterial ancestry, 803 of eubacterial ancestry and 138 of an unresolved prokaryotic ancestry.

The mosaic structure of the *S. cerevisiae* chaperone repertoire

Of the 69 known yeast molecular chaperones, 47 had homologs in prokaryotic genomes. These were classified based on their tree topology into 11 chaperones of archaebacterial ancestry and 26 chaperones of a eubacterial ancestry. The ancestry of the remaining ten chaperones could not be resolved from the data (Figure 1). The Hsp90 family in yeast includes two paralogs whose sequences are highly similar (96% identity at the amino acid level). Both paralogs are homologous to eubacterial htpG sequences exclusively, and hence the yeast Hsp90 is clearly of eubacterial origin. The prefoldin (PFD) chaperones transfer target proteins to the CCT system for further folding (Vainberg et al. 1998). The yeast genome encodes six PFD paralogs whose protein sequences are 15.2±3.8% identical. Three of the six PFDs have homologs in prokaryotic genomes, all of which are archaebacterial. The remaining three paralogs had no detectable homologs in prokaryotic genomes applying the sequence similarity threshold used in this study (>25% identical amino acids). This indicates that prefoldin is an archaebacterial contribution to eukaryotic genomes and the family further diversified within eukaryotes. All five small heat shock proteins (sHsp) were inferred to be of eubacterial ancestry. Hsp26 is homologous to eubacterial sequences only and the four paralogous genes Hsp31, Hsp32, Hsp33 and Sno4 clearly branch within the eubacterial clade, although homologs in halophilic and methanogenic archaebacteria were found as well. Members of the Hsp100 chaperone family (Clp) play a role in protein disaggregation (Parsell et al. 1994). Of the three Hsp100 proteins in yeast, one is localized in the mitochondria and two are cytosolic (van Dyck et al. 1998). The mitochondrial Clp protein Mcx1 was inferred to be of eubacterial origin. The cytosolic Hsp104 was inferred to have been derived from an archaebacterial AAA+ ATPase, while the second cytosolic Hsp78 is of ambiguous ancestry. The Hsp40 and Hsp70 families include chaperones with eubacterial as well as archaebacterial ancestry, although the majority of chaperones from these particular families are of eubacterial descent.

Eukaryotic genomes typically encode two chaperonin systems: the type I mitochondrial Hsp60/Hsp10 system (GroEL/ES-like) and the type II chaperonin (CCT-like). The type I chaperonin system is usually viewed as a eubacterial set of chaperones, however, it is also encoded in the genomes of several methanogenic and halophilic archaebacteria (e.g. Deppenmeier et al. 2002). The yeast Hsp60 branched in between a purely archaebacterial clade and a purely eubacterial clade. Consequently it was classified as of ambiguous prokaryotic ancestry. The co-chaperone Hsp10 is clearly of eubacterial origin. This classification fits well with its localization in the mitochondrion. The type II eukaryotic chaperonins comprise eight different protein subunits (Archibald et al. 1999; Valpuesta et al. 2002). These chaperones are usually viewed as archaebacterial, however, several Clostridia species encode type II chaperonins as well (Techtmann and Robb 2010; Williams et al., 2010). An archaebacterial ancestry was inferred for Tcp1 and a eubacterial origin was inferred for Cct4 and Cct8. The other five CCT genes were classified as ambiguous as they branch between clostridial and archaebacterial homologs.

Connectivity in the chaperone interaction network and protein ancestry

The chaperone–substrate interaction (CSI) network is based on a large-scale screening for proteins that interact with 63 of the 69 chaperones encoded in *S. cerevisae* (Gong et al. 2009). The CSI network contains 3,595 substrate proteins that interact with at least one chaperone and a total of 21,687 chaperones-substrate interactions. Interactions in the CSI network are unweighted and do not reflect their relative prevalence. We reduced the dataset to include only those chaperones and substrates for which prokaryotic ancestry could be determined. This network contained 37 chaperones and 1055 substrates. A total of 2691 interactions included in the network were classified into four classes based on the ancestry of both the chaperones and substrates (inset in Figure 2).

Substrates of archaebacterial ancestry interact with significantly more chaperones than eubacterial substrates (Wilcoxon-ranksum test, $p = 3.3\times10^{-04}$). The network connectivity pattern is not biased towards a higher number of interactions between chaperones and substrates of the same ancestry (χ^2 test; p = 0.32, inset in Fig. 2). This conclusion still holds when considering only substrates that interact with at least two chaperones or more (χ^2 test; p = 1.0). To further test for possible biases in the network connectivity pattern we examined the ratio of eubacterial to archaebacterial interaction partners for each chaperone and substrate, and tested for differences in the distributions of ratios in the two ancestry groups. We found no significant difference in the distributions of the chaperone ancestry ratio between archaeal and eubacterial substrates (Wilcoxon-ranksum test, $p = 0.99$), and no significant difference in the distributions of the substrate ancestry ratio between archaeal and eubacterial chaperones (Wilcoxon-ranksum test, $p = 0.081$). We further tested if any of the four chaperone–substrate ancestry combinations is enriched in the network by conducting a network randomization test with 10,000 randomization replicates (Figure 2). This analysis shows that none of the four interaction types is found at a frequency that is significantly different from the random expectation (at a false discovery rate (FDR) of 0.01).

Protein ancestry and protein function

Substrates in the network were further classified into two major functional categories according to their annotation in the Gene Ontology database (GO, Ashburner et al. 2000). Substrates whose annotation includes the terms “translation,” “transcription,” “DNA-dependent DNA replication”, or their subterms, were classified as proteins performing an informational function. The remaining substrates were classified as operational proteins (Rivera et al. 1998; Cotton and McInerney 2010). Combining the functional classification with prokaryotic ancestry reconstruction revealed that 69% of the 274 archaebacterial substrates and 14% of the 705 eubacterial substrates found in GO perform informational functions. Hence, substrates of archaebacterial origin are enriched for informational functions ($p<10^{-16}$, using χ^2 test), confirming the known correlations between prokaryotic ancestry and protein function (Esser et al. 2004; Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011; Alvarez-Ponce et al. 2013). As expected from the congruence between the ancestral and functional classifications, we found that informational substrates, like archaebacterial substrates, interact with a larger number of different chaperones than operational substrates (Wilcoxon-ranksum test, $p < 10^{-16}$).

Prokaryotic ancestry and protein physicochemical properties

A comparison of substrate physicochemical properties between the two ancestry groups revealed several significant differences. Eubacterial substrates were found to be longer on average, in agreement with previous studies (Alvarez-Ponce and McInerney 2011). In addition, eubacterial substrates are also enriched in hydrophobic and aromatic amino acids in comparison to archaebacterial substrates. Archaebacterial substrates are more conserved, more highly expressed and are encoded by higher proportions of preferred codons than eubacterial substrates (Figure 3). Biases in the three latter properties fit well with the known correlation among evolutionary rates, expression level and codon usage bias (Grantham et al. 1981; Sharp and Li 1987; Pál et al. 2001; Drummond et al. 2005; Pál et al. 2006). In addition, substrates of archaebacterial origin were enriched for positively charged amino acids as well as alanine, arginine, glutamate, lysine and valine. On the other hand, substrates of eubacterial origin are significantly enriched in cysteine, histidine, isoleucine, leucine, phenyl-alanine, proline, serine threonine, tryptophane and tyrosine (Figure 3). Most of the above differences in substrate physicochemical properties are observed also when contrasting informational and operational proteins, as expected from the congruence between the ancestral and functional classifications.

Discussion

Our evolutionary reconstruction of the ancestry of chaperones involved in the yeast protein folding pathway reveals that chaperones of different descent are utilized in a coordinated fashion to fold common substrates. For example, the Hsp40/Hsp70 system in yeast comprises a total of 21 Hsp40 and 14 Hsp70 genes from diverse origins including archaebacterial, eubacterial, and eukaryotic-specific proteins (ESPs). Interestingly, the Hsp40 family, with eleven ESPs, has diversified within eukaryotes to a larger extent in comparison to the Hsp70 family that includes only one ESP. The difference between the two families can be explained by their mode of function. Chaperones of the Hsp40 family are the drivers of Hsp70 substrate activity and specificity (Cyr and Douglas 1994, Kampinga and Craig 2010). Thus, the diversification of Hsp40 family within eukaryotes probably enabled the whole Hsp40-Hsp70 system to increase its operational potential. A mosaic of ancestries is observed in all chaperone families that are present in both archaebacteria and eubacteria. It is noteworthy that in contrast to cytosolic chaperones, yeast chaperones that are localized in the mitochondria are an exception. All mitochondrial chaperones that could be classified by their tree topology are inferred to be of eubacterial ancestry, underlining the role of the mitochondrion as a functional eubacterial unit within the eukaryotic cell (Esser et al. 2004).

Previous studies showed that there is a significant preference for proteins to interact with partners of the same ancestry rather than across the archaebacterial-eubacterial divide (Alvarez-Ponce and McInerney 2011). Such preference can be expected if the proteins participating in specific cellular pathways are usually of a single ancestry. Since protein connectivity is higher within pathways than across pathways, common ancestry of pathway proteins will result in an overall trend for same ancestry interactions. Thus, same ancestry preference, while demonstrable on average, may still be violated when considering specific systems. Our results suggest that the general trend does not hold for the chaperone–substrate interactions network, where no preference for interaction of chaperones and substrates of the same ancestry could be observed. This indicates that the protein folding pathways have been reorganized and integrated to a larger extent in comparison to the overall protein-protein interactions within the cell. The only difference correlated with ancestry in the chaperone–substrate interaction network is to be found in the tendency of archaebacterial substrates to have a larger number of chaperone interactions than eubacterial substrates , in agreement with the general trend observed before (Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011).

What makes molecular chaperones a class of proteins that is more amenable to integration? Chaperones are highly versatile proteins that increase the probability of their substrates to attain a functional conformation and by that can contribute significantly to the organismal fitness. Chaperones are essential in both prokaryotic domains (Hartl and Hayer-Hartl 2002; Calloni et al. 2012); hence, at the very origin of the eukaryotic cytosol, there was an absolute need for chaperones of both ancestries to assist in the folding of their respective substrates. Yet, similar chaperones may have similar substrate specificity and interact with a similar set of proteins, so eubacterial and archaeal chaperones might have had overlapping substrate sets already at the beginning, though not in a regulatory coordinated context. A non-specific interaction pattern allows chaperones to acquire new clients without the need for intensive sequence modification or adaptation, and the evolution of a completely integrated system is expected to have included also the co-expression of substrates and their chaperones as well as optimizing the competitive binding of substrates and their dedicated chaperones. The effects of combining two ancestral chaperone systems may have conferred an even larger fitness benefit than was possible by either of the ancestral systems on its own. Nonetheless, retaining two chaperone systems would have entailed an additional energetic cost for the cell as chaperone-mediated folding is expensive in terms of ATP usage. In the context of eukaryogenesis, this would not have posed an insurmountable problem, since the formation of mitochondria as an intracellular organelle resulted in a dramatic increase in the available energy for all cellular processes (Lane and Martin 2010). Nevertheless, energetic considerations might still play a role in the evolution of chaperone–substrate interactions (Bogumil and Dagan 2012).

In summary, in contrast with other proteins that still show a tendency to form network communities reflecting their ancestries, molecular chaperones have been able to cross the divide between the ancestral prokaryotic domains. The central role of chaperone-assisted folding in maintaining cellular fitness is reflected in the high degree of integration of an archaebacterial and a eubacterial chaperone systems into one at the origin of Eukaryotes.

Methods

Data

Yeast protein sequences, amino acid usage data, functional assignments, chromosomal locations, frequencies of optimal codons, codon adaptation indexex (CAI), gravy scores (hydropathy index), and aromaticity scores were downloaded from the *Saccharomyces* Genome Database (SGD) (Cherry et al. 1998). Chaperone-protein interaction data were obtained from Gong et al. (2009). The secondary structure of all proteins was inferred using PsiPred (Jones 1999), applying a threshold of 70% for the calculation of secondary structure probability. Quantitative protein expression data were obtained from Ghaemmaghami et al. (2003). The mRNA levels data were obtained from Wang et al. (2002). For the statistical analysis of protein expression levels, natural log-transformation was applied. Proteins for which expression levels were not available (107 in total) or with zero expression level (1,665 proteins) were excluded from the analysis. All statistical analyses were performed using the MatLab© Statistics toolbox.

Evolutionary Rate

Positional orthology assignments among 20 fungal genomes were obtained from Wapinski et al. (2007). Proteins lacking orthologs in any genome (282 in total) were excluded from the analysis. Multiple sequence alignments of all yeast open reading frames with orthologous sequences were generated with MAFFT (Katoh et al. 2005). Phylogenetic trees were reconstructed with PhyML v3.0_360-500M (Guindon and Gascuel 2003) using the best-fit model as inferred by ProtTest 3 (Darriba et al. 2011) according to the Akaike information criterion measure (Akaike 1974). Distances from the *S. cerevisiae* proteins to their orthologs were calculated as the sum of branch lengths. To calculate the relative amino acid substitution rates of substrates, the distances to the 20 proteomes were first Z-transformed separately and then averaged over all orthologs (Bogumil and Dagan 2010).

Reconstruction of prokaryotic ancestries

We classified each of the 5880 yeast protein-coding genes into archaebacterial, eubacterial, ambiguous prokaryotic ancestry, or eukaryote-specific, based on its phylogenetic affinities. Each yeast protein sequence was used as query in a homology search against a database containing the proteomes of 82 archaebacteria and 1074 eubacteria (3,792,506 proteins in total). Homology searches were carried out using PSI-BLAST (Altschul et al. 1997) without filtering. Global pairwise alignments of BLAST-hits were calculated using the EMBOSS package (Needleman and Wunsch 1970, Rice et al. 2000). Prokaryotic sequences with less than 25%

identity were considered as having no significant similarity to the particular yeast query. Of the yeast genes, 161 had significant similarity to archaebacterial sequences exclusively (and were thus classified as being of archaebacterial ancestry), 383 had significant similarity to eubacterial sequences only (and were thus deemed as eubacterium-derived), and 686 had homologs in both prokaryotic domains. The remaining genes had no detectable prokaryotic homologs at the specified thresholds, and were thus considered eukaryote-specific.

In order to ascertain the ancestry of the 686 yeast genes with both archaebacterial and eubacterial homologs, we conducted a phylogenetic analysis. For each of these genes, a multiple sequence alignment including the 15 best BLAST hits from each prokaryotic domain was generated using MAFFT v6.843b (Katoh and Toh 2008) and the quality of the alignment was tested with Guidance (Penn et al. 2010). In order to be conservative in our analysis, columns with a confidence score < 0.93 were removed. Phylogenetic trees were reconstructed with PhyML v3.0_360-500M (Guindon and Gascuel 2003) using the best-fit model as inferred by ProtTest 3 (Darriba et al. 2011) according to the Akaike information criterion (Akaike 1974).

We next rooted each tree on the branch that maximized the separation of archaebacterial and eubacterial sequences. The internal branch yielding the maximum ratio of archaebacteria to eubacteria content in the resulting clades was determined with the MRP function implemented in CLANN 3.2.2 (Creevey and McInerney 2005) using Spearman's rank correlation coefficient. The yeast gene was classified as of eubacterial or archaebacterial ancestry depending on the clade within which it branched. Yeast genes were considered of ambiguous ancestry if no branch yielded a clear separation into an archaebacterial and eubacterial clades, if multiple branches separated the archaebacterial and eubacterial sequences equally well and resulted in conflicting ancestry assignments, or if the yeast gene branched between the archaebacterial and eubacterial clades. In such ambiguous cases, we repeated the analysis with a larger sample of homologous sequences, first with the 30 best BLAST hits from each domain, and if still ambiguous, with the 45 best BLAST hits from each domain. This analysis shifted 125 genes from the ambiguous to the unambiguous class. Of the 686 yeast genes with both archaebacterial and eubacterial homologs, 128 were classified as of archaebacterial ancestry, 420 as of eubacterial ancestry and 138 as ambiguous.

In total, we inferred 289 proteins to be of archaebacterial ancestry, 803 of eubacterial ancestry and 138 proteins with an unresolvable prokaryotic ancestry. The remaining yeast proteins did not show significant similarity with any prokaryotic protein.

Network randomization

Randomization of the CSI network was carried out using the switching methodology (Stone and Roberts 1990; Artzy-Randrup and Stone 2005), implemented in an in-house MatLab script.

Acknowledgements

We thank Robin Koch, Ovidiu Popa, and Thorsten Thiergart for their helpful comments on the manuscript and acknowledge support from the European Research Council (grant No. 281357).

Figure Legends

Figure 1. An illustration of reconstructed ancestries for yeast major chaperone families. Archaebacterial ancestry is shown in red, and eubacterial ancestry in blue. Chaperones with ambiguous ancestry or no homology to prokaryotic proteins are colored in purple and grey, respectively. Here we use the same model for all members of the same family. We note however that paralogs may deviate in their protein structures. Molecule plots were generated using the PyMOL Molecular Graphics System, version 1.5.0.4 (Schrödinger, LLC).

Figure 2. Prokaryotic origin and connectivity distribution. Asterisks indicate the observed percentage of edges in the network while the bars show the mean expected frequency from randomization simulations. The lines indicate the 1-99 percentile range. Abbreviations: A-Archaebacterial, B-Eubacterial; upper-case indicate chaperones and lower-case indicate substrates.

Figure 3. Differences in protein physicochemical properties between substrates of eubacterial and archaebacterial origin. Bars indicate a significantly higher value for the corresponding group of the particular property. The bar length corresponds to a $\log_{10}$ transformation of the p-value in a two-tailed Kolmogorov-Smirnov test.

References

Akaike H. 1974. A new look at the statistical model identification. Automatic Control, IEEE T. Automat. Contr. 19:716–723.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Alvarez-Ponce D, McInerney JO. 2011. The Human Genome Retains Relics of Its Prokaryotic Ancestry: Human Genes of Archaebacterial and Eubacterial Origin Exhibit Remarkable Differences. Genome Biol. Evol. 3:782–790.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc. Natl. Acad. Sci. U.S.A. 110:1594-1603.

Archibald JM, Logsdon JM, Doolittle WF. 1999. Recurrent paralogy in the evolution of archaeal chaperonins. Curr. Biol. 9:1053–1056.

Artzy-Randrup Y and Stone L. 2005. Generating uniformly distributed random networks. Phys. Rev. E 72:056708.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25:25–29.

Bogumil D, Dagan T. 2010. Chaperonin-dependent accelerated substitution rates in prokaryotes. Genome Biol. Evol. 2:602–608.

Bogumil D, Landan G, Ilhan J, Dagan T. 2012. Chaperones Divide Yeast Proteins into Classes of Expression Level and Evolutionary Rate. Genome Biol. Evol. 4:618−625.

Bogumil D, Dagan T. 2012. Cumulative Impact of Chaperone-Mediated Folding on Genome Evolution. Biochemistry 51:9941–9953.

Calloni G, Chen T, Schermann SM, Chang H-C, Genevaux P, Agostini F, Tartaglia GG, Hayer-Hartl M, Hartl FU. 2012. DnaK Functions as a Central Hub in the E.&nbsp;coli Chaperone Network. Cell Rep. 1:251–264.

Cherry J, Adler C, Ball C, et al. 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26:73–79.

Cotton JA, McInerney JO. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc. Natl. Acad. Sci. U.S.A. 107:17252–17255.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 105:20356–20361.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21:390–392.

Cyr DM, Douglas MG. 1994. Differential regulation of Hsp70 subfamilies by the eukaryotic DnaJ homologue YDJ1. J. Biol. Chem. 269:9798–9804.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Deppenmeier U, Johann A, Hartsch T, et al. 2002. The genome of Methanosarcina mazei: Evidence for lateral gene transfer between bacteria and archaea. J. Mol. Microbiol. Biotechnol. 4:453–461.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U.S.A. 102:14338–14343.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature 440:623–630.

Esser C, Ahmadinejad N, Wiegand C, et al. 2004. A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. 21:1643–1660.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. GroEL buffers against deleterious mutations. Nature 417:398

Gabaldón T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. Science 301:609–609.

Ghaemmaghami S, Huh W, Bower K, Howson R, Belle A, Dephoure N, O'Shea E, Weissman J. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.

Gong Y, Kakihara Y, Krogan N, Greenblatt J, Emili A, Zhang Z, Houry WA. 2009. An atlas of chaperone-protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. Molecular Systems Biology 5:1–14.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9:43−74.

Gray MW. 1999. Mitochondrial Evolution. Science 283:1476–1481.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hartl FU, Hayer-Hartl M. 2009. Converging concepts of protein folding in vitro and in vivo. Nat. Struct. Mol. Biol. 16:574–581.

Hartl FU. 2002. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. Science 295:1852–1858.

Jones DT. 1999. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. J. Mol. Biol. 292:195–202.

Kampinga HH, Craig EA. 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. Nat. Rev. Mol. Cell Biol. 11:579–592.

Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9:286–298.

Lane N. 2009. Life ascending: the ten greatest inventions of evolution. London: Profile Books. 344 p.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929–934.

Large AT, Goldberg MD, Lund PA. 2009. Chaperones and protein folding in the archaea. Biochem. Soc. Trans 37:46.

Large AT, Lund PA. 2009. Archaeal chaperonins. Front. Biosci. 14:1304–1324.

Macario AJ, Dugan CB, Clarens M, Conway de Macario E. 1993. dnaJ in Archaea. Nucleic Acids Res. 21:2773.

Macario AJ, Dugan CB, Conway de Macario E. 1991. A dnaK homolog in the archaebacterium Methanosarcina mazei S6. Gene 108:133–137.

Maisnier-Patin S, Roth JR, Fredriksson Å, Nyström T, Berg OG, Andersson DI. 2005. Genomic buffering mitigates the effects of deleterious mutations in bacteria. Nat. Genet. 37:1376–1379.

Martin W, Herrmann R. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol. 118:9–17.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.

Martin W. 2003. Gene transfer from organelles to the nucleus: frequent and in big chunks. Proc. Natl. Acad. Sci. U.S.A. 100:8612–8614.

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. 48:443–453.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. . Proc. Natl. Acad. Sci. U.S.A. 109:20537–20542.

Parsell DA, Kowal AS, Singer MA, Lindquist S. 1994. Protein disaggregation mediated by heat-shock protein Hsp104. Nature 372:475–478.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nature *Nat.* Rev. Genet. 7, 337–348.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res. 38:W23–W28.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes. Mol. Biol. Evol. 24:1752–1760.

Queitsch C, Sangster T, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. Nature 417:618–624.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. U.S.A. 95:6239–6244.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431:152–155.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. Nature 396:336–342.

Sagan L. 1967. On the origin of mitosing cells. J. Theoret. Biol. 14:255–274.

Sharp PM, Li WH. 1987. The Codon Adaptation Index: A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281−1295.

Stone L and Roberts A. 1990. The checkerboard score and species distribution. Oecologia 85: 74-79.

Techtmann SM, Robb FT. 2010. Archaeal-like chaperonins in bacteria. Proc. Natl. Acad. Sci. U.S.A. 107:20269− 20274.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin. Genome Biol. Evol. 4:466–485.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5, 123–135.

Vainberg I, Lewis S, Rommelaere H, Ampe C, Vandekerckhove J, Klein H, Cowan N. 1998. Prefoldin, a chaperone that delivers unfolded proteins to cytosolic chaperonin. Cell 93:863–873.

Valpuesta J, Martin-Benito J, Gomez-Puertas P, Carrascosa J, Willison K. 2002. Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT. FEBS Lett. 529:11–16.

van Dyck L, Dembowski M, Neupert W, Langer T. 1998. Mcx1p, a ClpX homologue in mitochondria of Saccharomyces cerevisiae. FEBS Lett. 438:250–254.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. U.S.A. 99:5860–5865.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61.

Williams TA, Fares MA. 2010. The Effect of Chaperonin Buffering on Protein Evolution. Genome Biol. Evol. 2:609–619.

Williams TA, Codoñer FM, Toft C, Fares MA. 2010. Two chaperonin systems in bacterial genomes with distinct ecological roles. Trends Genet. 26:47–51.

Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. P. Roy. Soc. B-Biol. Sci. 279, 4870–4879. 279:4870–4879.

Young J, Agashe V, Siegers K, Hartl F. 2004. Pathways of chaperone-mediated protein folding in the cytosol. Nat. Rev. Mol. Cell Biol. 5:781–791.

Figure 1

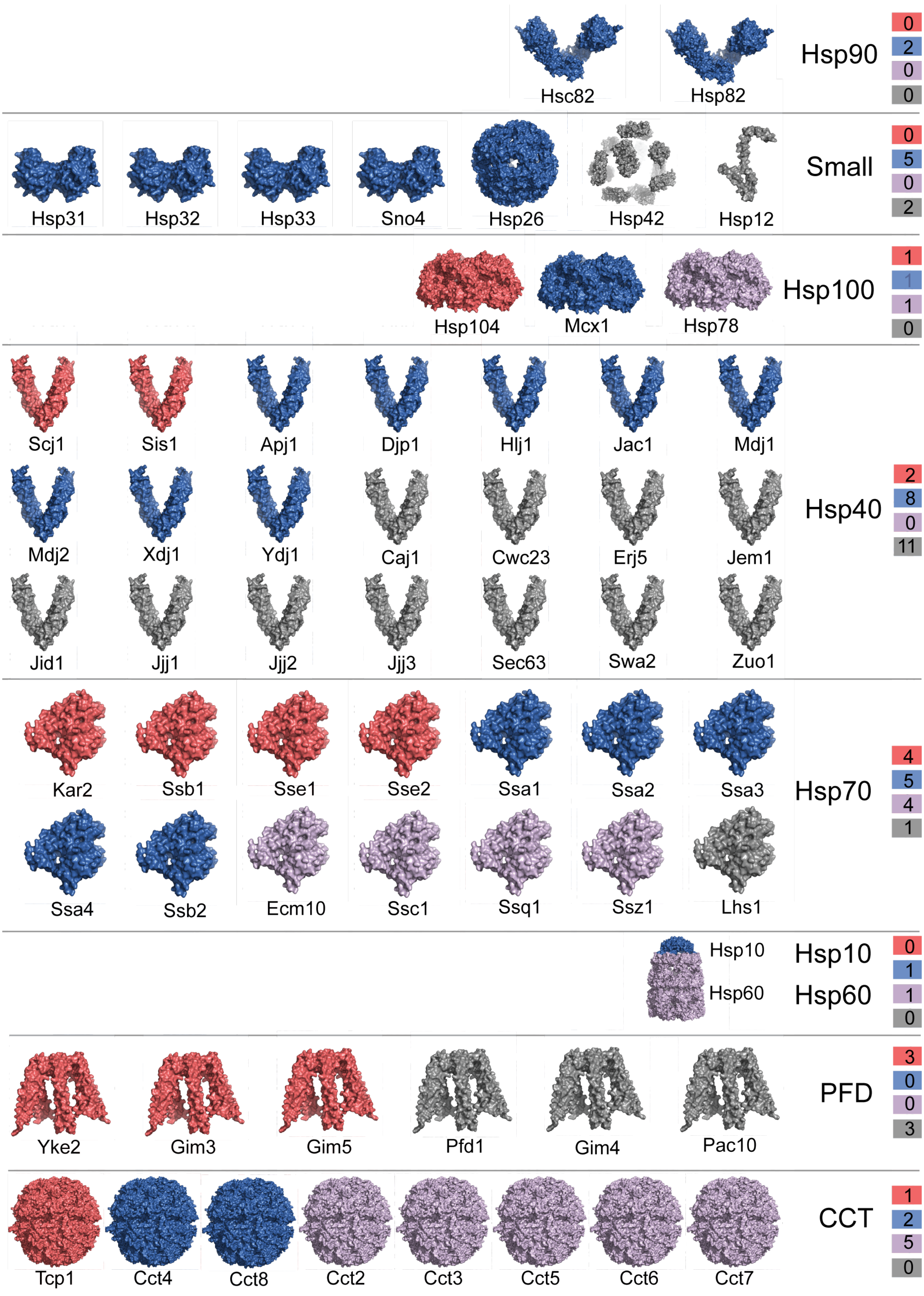

Hsc82
Hsp82
Hsp90
0
2
0
0
Hsp31
Hsp32
Hsp33
Sno4
Hsp26
Hsp42
Hsp12
Small
0
5
0
2
Hsp104
Mcx1
Hsp78
Hsp100
1
1
1
0
Scj1
Sis1
Apj1
Djp1
Hlj1
Jac1
Mdj1
Mdj2
Xdj1
Ydj1
Caj1
Cwc23
Erj5
Jem1
Jid1
Jjj1
Jjj2
Jjj3
Sec63
Swa2
Zuo1
Hsp40
2
8
0
11
Kar2
Ssb1
Sse1
Sse2
Ssa1
Ssa2
Ssa3
Ssa4
Ssb2
Ecm10
Ssc1
Ssq1
Ssz1
Lhs1
Hsp70
4
5
4
1
Hsp10
Hsp60
Hsp10
Hsp60
0
1
1
0
Yke2
Gim3
Gim5
Pfd1
Gim4
Pac10
PFD
3
0
0
3
Tcp1
Cct4
Cct8
Cct2
Cct3
Cct5
Cct6
Cct7
CCT
1
2
5
0

Figure 2

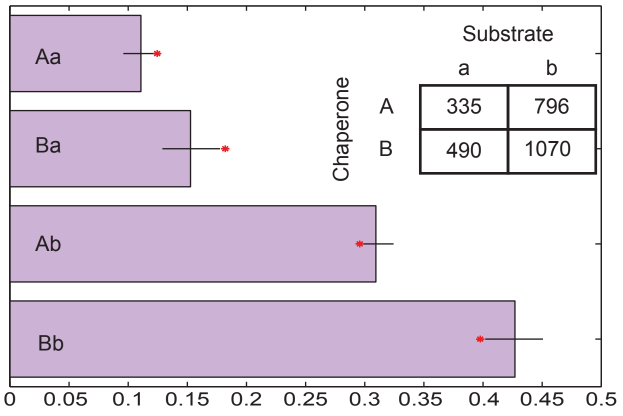
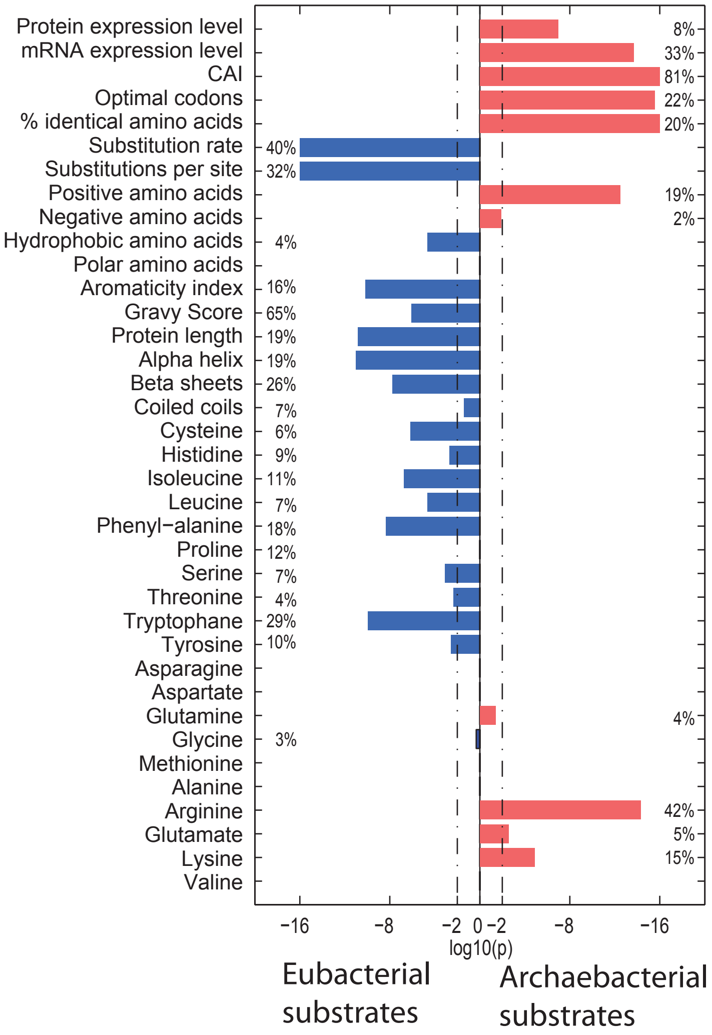


Figure 3



6 Acknowledgements

Thank you…

First and foremost I would like to thank my supervisor Prof. Dr. Tal Dagan for the opportunity to work on the projects that now comprise this thesis. Her time, ideas, advice, patience and funding not only made this possible but also an encouraging and productive experience.

I would like to thank Prof. Dr. William F. Martin for his support from the beginning of my undergraduate studies to the end of my Phd. It was a great time, not only science-wise I learned a lot.

Furhtermore I want to thank Dr. Giddy Landan for many fruitful discussions on the fields of molecular evolution and statistics, new perspectives on almost any form of weird art and one or the other cigarette at the right time.

All the colleagues, friends and people I met along the way. Without you it would have been a lot less enjoyable. Thank you: Ovidiu Popa, Robin Koch, Julia Weissenbach, Judith Ilhan and the past and present team of the Dagan Group. Kathrin Hoffmann, Dr. Mayo Röttger, Christian Wöhle, Thorsten Tiergart, Dr. Shijulal Nelson-Sathi , Dr. Verena Zimorski, Dr. Peter Major, Dr. Houda El-Haddad, Thorsten Klösges, Dr. Christian Esser and the former teammates in the molevol group in Düsseldorf.

My family who always supported me and my countless stupid and few wise decisions in my life. Thank you for making this possible.

And finally I want to thank Michaela Hartmann for being able to live with me. Without your continuous support and inspiration all this could not have been done.