On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life

Mathieu E. Rebeaud b,1, Saurav Mallik b,1, Pierre Goloubinoff a,2 and Dan S. Tawfik b,3

1Department of Plant Molecular Biology, Faculty of Biology and Medicine, University of Lausanne, CH-1015 Lausanne, Switzerland; and 2Department of Biomolecular Sciences, The Weizmann Institute of Science, 7610001 Rehovot, Israel

Edited by Lila M. Giersch, University of Massachusetts, Amherst, MA, and approved April 2, 2021 (received for review October 6, 2020)

Across the Tree of Life (ToL), the complexity of proteomes varies widely. Our systematic analysis depicts that from the simplest archaea to mammals, the total number of proteins per proteome expanded ∼200-fold. Individual proteins also became larger, and multidomain proteins expanded ∼50-fold. Apart from duplication and divergence of existing proteins, completely new proteins were born. Along the ToL, the number of different folds expanded ∼5-fold and fold combinations ∼20-fold. Proteins prone to misfolding and aggregation, such as repeat and beta-rich proteins, proliferated ∼600-fold and, accordingly, proteins predicted as aggregation-prone became 6-fold more frequent in mammalian compared with bacterial proteomes. To control the quality of these expanding proteomes, core chaperones, ranging from heat shock proteins 20 (HSP20s) that prevent aggregation to HSP60, HSP70, HSP90, and HSP100 acting as adenosine triphosphate (ATP)-fueled unfolding and refolding machines, also evolved. However, these core chaperones were already available in prokaryotes, and they comprise ∼0.3% of all genes from archaea to mammals. This challenge—roughly the same number of core chaperones supporting a massive expansion of proteomes—was met by 1) elevation of messenger RNA (mRNA) and protein abundances of the ancient generalist core chaperones in the cell, and 2) continuous emergence of new substrate-binding and nucleotide-exchange factor cochaperones that function cooperatively with core chaperones as a network.

Tree of Life | expansion of proteomes | core chaperones | cochaperones | chaperone network

All cellular life is thought to have stemmed from the last universal common ancestor (LUCA) (1, 2), that emerged more than 3.6 billion y ago. Two major kingdoms of life diverged from LUCA: bacteria and archaea, which about 2 billion y later merged into the eukaryotes (3). Since the beginning of biological evolution, life’s volume has increased on a grand scale: The average size of individual cells has increased ∼100-fold from prokaryotes to eukaryotes (4), the number of cell types has increased ∼200-fold from unicellular eukaryotes to humans (5), and average body size has increased ∼5,000-fold from the simplest sponges to blue whales (6).

This expansion in organismal complexity and variability was accompanied by an expansion in life’s molecular workforce, proteomes in particular, which in turn presented a challenge of reaching and maintaining properly folded and functional proteomes. Most proteins must fold to their native structure in order to function, and their folding is largely imprinted in their primary amino acid sequence (7–9). However, many proteins, and especially large multidomain polypeptides, or certain protein types such as all-beta or repeat proteins, tend to misfold and aggregate into inactive species that may also be toxic (10). Life met this challenge by evolving molecular chaperones that can minimize protein misfolding and aggregation, even under stressful out-of-equilibrium conditions favoring aggregation (11, 12). Chaperones can be broadly divided into core and cochaperones. Core chaperones can function on their own, and include ATPases heat shock protein 60 (HSP60), HSP70, HSP100, and HSP90 and the adenine triphosphate (ATP)-independent HSP20. The basal protein holding, unfolding, and refolding activities of the core chaperones are facilitated and modulated by a range of cochaperones such as J-domain proteins (13–15).

Starting from LUCA, as proteomes expanded, so did the core chaperones and their respective cochaperones. Indeed, chaperones have been shown to facilitate the acquisition of destabilizing mutations and thereby accelerate protein evolution (16–18). However, the coexpansion of proteomes and of chaperones, underscoring a critical balance between evolutionary innovation and foldability, remains largely unexplored. We thus embarked on a systematic bioinformatics analysis that explores the evolution of both proteomes and chaperones, and of both core and their auxiliary cochaperones, along the Tree of Life.

Results

A Tree of Life Analysis of the Expansion of Proteomes and Chaperones.

We aimed to explore, systematically, across the Tree of Life (ToL) the expansion of proteomes and compare it with the chaperone composition and level. To this end, we collected proteome sequences from representative organisms belonging to all the major bacterial, archaeal, and eukaryotic clades and constructed a ToL. This expansion in organismal complexity and variability was accompanied by a massive expansion of the protein universe. Compared with simple prokaryotes that harbor thousands of proteins, plants and animals harbor hundreds of thousands of proteins that are also longer, multidomain, and comprise a variety of folds and fold combinations, repeated segments, and beta-rich architectures that make them prone to misfolding and aggregation. Surprisingly, the relative representation of core chaperones, those dedicated to maintaining the folding quality of these increasingly complex proteomes, did not change from prokaryotic to mammalian genomes. To reconcile the expanding proteomes, core chaperones have rather increased in cellular abundance and evolved to function cooperatively as a network, combined with their supporting workforce, the cochaperones.

Significance

Across the Tree of Life, life’s phenotypic diversity has been accompanied by a massive expansion of the protein universe. How-
The overall topology of our tree was borrowed from TimeTree (19) and we also adhered to their order of divergence, which is based on molecular dating and geological records. TimeTree also provides putative dates of emergence and these are provided as branch lengths, yet because our analysis is primarily comparative, branch lengths were only used here as a graphical aid (Fig. 1A).

The Tree of Life begins with LUCA at the root (Fig. 1A). The edges of the ToL represent the extant three kingdoms—archaea, plotted throughout in black; bacteria, plotted in blue; and eukaryotes. The latter emerged by endosymbiosis of an Alphaproteobacterium and an Asgard-like archaeon (20, 21). The emergence of green algae and subsequently of plants occurred with a secondary endosymbiosis of a Cyanobacterium into a nonphotosynthetic eukaryote (22). The major eukaryotic clades therefore comprised unicellular, early-diverging eukaryotes (in orange), fungi (gray), plants (green), Metazoa (invertebrate animals; red), and Chordata (vertebrate animals; wine). Overall, our analysis was based on comparing the proteomes of 188 representative organisms, covering 56 major clades of bacteria, archaea, and eukaryote (Dataset S1). The various proteome parameters analyzed below were initially derived for each representative organism in the core tree. The representative organisms of each clade were then pulled together to calculate the clade average and the SD for this average. The clade average values were subsequently plotted using the order of divergence for the x axis. Accordingly, these plots also broadly divide into prokaryotes (the left part) and eukaryotes (the right part), and the latter’s right edge comprises Chordata including Mammalia (Fig. 1B and the following figures).

### Expansion of Proteome Size

Initially, we scrutinized the expansion of proteome size by examining 1) the total number of proteins per proteome in a given clade; 2) the median protein length; and 3) the number of multidomain proteins in the proteome. The clade average values of these three parameters are plotted in Fig. 1B–D, with colors of points matching the branch colors in Fig. 1A.

**The total number of proteins per proteome expanded ~200-fold.** Proteomes that comprise a larger number of proteins unavoidably present a greater challenge for their protein quality control chaperone machinery. To examine the expansion in the number of proteins per proteome, proteome sequences of the 188 representative organisms were obtained. Across the ToL, the number of proteins per proteome expanded roughly 200-fold (Fig. 1B) from ~700 proteins in the simplest free-living DPANN (23) archaea to

---

**Fig. 1.** Expansion of proteome size across the Tree of Life. (A) The ToL used in this study. Leaves represent extant phylogenetic clades, while internal nodes represent their presumed ancestors. Branch lengths are in million years, as available from TimeTree (19), and they refer to the relative order of divergence of the corresponding clades rather than the absolute dates of their emergence. The major phylogenetic groups in the tree (Bacteria, Archaea, unicellular eukaryotes, plants, Fungi, Metazoa, and Chordata) are highlighted in different colors. Vertical arrows highlight the two major endosymbiosis events: the Alphaproteobacterial origin of mitochondria and the cyanobacterial origin of plastids. LAA, last archaeal ancestor; LBA, last bacterial ancestor. (B–D) The average per-clade values for various proteome size parameters (y axis, in log scale) are plotted against their order of divergence (along the x axis in linear scale). These parameters include proteome size (B), median protein length (C), and multidomain proteins in the proteome (D). In these scatterplots, the colors of data points represent their major phylogenetic group in the tree (A). Error bars represent the clade SD (no error bars relate to clades comprising only one representative organism). The lines were derived by a fit to an exponential equation, and are provided merely as visual guides. Prokaryotic and eukaryotic organisms are separated by a dashed line in B.
Proteome size is similar across prokaryotes, on the order of 3,000 proteins, although the smallest proteomes belong to the earliest-diverging free-living DPANN archaea and Aquificae bacteria (23, 24) (not counting parasites and symbionts). Eukaryote proteomes are substantially larger and, considering free-living organisms only, the smallest eukaryotic proteomes harboring ~10,000 proteins belong to Amoebozoa, one of the earliest-diverging eukaryotes (25, 26). Land plants and metazoaos comprise hundreds of thousands of proteins per proteome. However, as described later, this dramatic increase in the number of proteins in eukaryotes occurred not only by duplication of preexisting proteins but also by the emergence of completely new domains and folds.

**The median protein length increased ~2-fold.** The longer the polypeptides are the more prone they are to misfold and aggregate instead of readily reaching their native functional state (27, 28). Analyzing the lengths of all proteins in each representative proteome (SI Appendix), we found that compared with ~250 residues across prokaryotes, median protein length increased about 2-fold in multicellular eukaryotes (Fig. 1C), with ~400 residues in plants and ~500 residues in Chordata (Dataset S2). Longer proteins were found primarily in multicellular eukaryotes (average lengths of the top 10% largest proteins were roughly 1,300 residues in plants, ~1,500 residues in metazoans, and ~2,150 residues in mammals). The longest polypeptides in mammals are predominantly muscle proteins, including different variants of titin (∼34,000 residues) or adhesins (>5,000 residues).

There are different ways by which proteins can increase in size. First, the domains themselves can grow larger by decorating an ancestral core domain with additional segments. Second, the fusion of multiple domains can result in a larger multidomain protein. Third, domain-flanking regions (C- and N-terminal segments, and interdomain linkers), that are typically disordered, can expand. A systematic analysis of 38 distinct folds that are conserved across the ToL (including parasites and symbionts) showed that lengths of individual domains increased mildly, nearly 1.5-fold, across the ToL (SI Appendix, Fig. S1A). The expansion of domain-flanking segments was also modest, nearly 3-fold, from prokaryotes to multicellular eukaryotes (SI Appendix, Fig. S1B). Indeed, as elaborated below, length expansion primarily stemmed from the increase in the fraction of multidomain proteins.

**Multidomain proteins expanded ~50-fold.** Multidomain proteins are inherently more prone to misfolding and aggregation than single-domain proteins, and may therefore demand more chaperone holding-unfolding-refolding action (29–31). To examine their expansion, domain annotations of all proteins in the 188 representative organisms were obtained from Pfam (32). Across the ToL, multidomain proteins comprising ≥3 Pfam-annotated domains have expanded ~50-fold (Fig. 1D), from ~100 proteins per proteome in prokaryotes to ~5,000 in plants and animals (Dataset S2). Further, multidomain proteins have expanded beyond the expansion of proteome size, to become nearly 3-fold more frequent in eukaryotic proteomes compared with prokaryotes (SI Appendix, Fig. S1C) with the expected corresponding shrinkage of proteins comprising one or two domains (SI Appendix, Fig. S1C and D). This expansion of multidomain proteins occurred not only by duplication, namely by amplifying preexisting multidomain proteins, but also via the emergence of new combinations. To assess the latter, we examined the number of unique combinations of domains per proteome, with domains being assigned by ECOD X-groups. It appeared that new domain combinations arose throughout evolution and, from prokaryotes to eukaryotes, the number of unique combinations per proteome increased ~20-fold (Fig. 2B), from ~100 combinations in bacteria to ~2,000 combinations in Chordata (Datasets S4 and S5).

**New fold combinations expanded ~20-fold.** As shown above, multidomain proteins expanded nearly 3-fold (their fraction out to the total number of proteins) alongside a parallel shrinkage of proteins comprising one or two domains (SI Appendix, Fig. S1C and D). This expansion of multidomain proteins occurred not only by duplication, namely by amplifying preexisting multidomain proteins, but also via the emergence of new combinations. To assess the diversity of new folds and fold combinations likely impose an additional burden on the chaperone machinery. We thus analyzed additional proteome parameters that represent expansion by innovation, rather than by mere duplication and divergence, as detailed below.

**Fold types expanded ~5-fold.** To assess the emergence of new folds, we used evolutionary classification of domains (ECOD)—a hierarchical classification of protein folds that uses both sequence and structural similarities and clusters all domains with known structures into independently evolved lineages, termed “X-groups” (33). For each representative organism, the Pfam-annotated domains were mapped to ECOD X-groups (SI Appendix). We found that from prokaryotes to eukaryotes, the number of unique folds (i.e., unique ECOD X-groups) per proteome expanded about 5-fold (Fig. 2A), from ~150 in prokaryotes to ~450 in metazoans (Dataset S3).

**All-beta and beta-rich folds expanded up to ~600-fold.** Proteins that are beta-rich are known to be prone to misfolding and aggregation (34). In the simplest free-living bacteria and archaea, proteins comprising the ancient all-alpha and alpha-beta architectures are the most frequent. Remarkably, upon the emergence of eukaryotes, and in metazoans especially, all-beta or beta-rich architectures (beta superfold) expanded massively, nearly 600-fold (Fig. 2C). Beta-rich proteins, in proportion to the total number of proteins, became nearly 6-fold more frequent in mammalian proteomes, as compared with bacteria and archaea (SI Appendix, Fig. 2S and Dataset S5). The immunoglobulin fold had a major contribution to this expansion, owing to its diverse roles in immunity, multicellularity, and signaling (35).

**Repeat sequences expanded ~700-fold.** Proteins comprising tandem repeats of nearly identical sequences emerge readily yet are prone to misfolding and aggregation. We identified proteins with repeated sequences of the size of a single “foldon” unit, ~20 amino acids (aa) (9), with ≥90% sequence similarity (Dataset S7). Most of the early-diverging archaea and bacteria do not possess repeat proteins. Indeed, repeat proteins appear in more recently diverged prokaryotes and foremost in eukaryotes (a similar trend was described in ref. 36). Indeed, metazoan proteomes contain large proteins with long repeated segments, for example, *Drosophila* Ank2p (21 Ankyrin repeats, in total 836 residues) or human Dmbt1p (11 cysteine-rich repeats of a total 1,419-residue length). The cumulative length of repeat sequences in metazoan proteomes can be up to 100,000 residues. Overall, from prokaryotes to eukaryotes, a 700-fold expansion (Fig. 2D) of repeated sequences was observed along the ToL (SI Appendix, Fig. S3). Repeated sequences also expanded beyond the expansion of proteome size—the percentage of total proteome length that comprises repeats increased nearly 7-fold from prokaryotes to metazoans (Fig. 2D).

**Proteome Expansion by Innovation.** Most proteins emerge by duplication and divergence of a preexisting protein. The outcome is paralogous proteins with the same overall fold and domain arrangement (for multidomain proteins). Thus, duplication and “local” divergence (point mutations and short insertions or deletions) certainly increase proteome size (the total number of proteins) but do not dramatically change proteome composition or complexity. The latter relates primarily to the birth of completely new proteins possessing new folds, and to the emergence of multidomain proteins with new fold combinations. Additions of new folds and fold combinations likely impose an additional burden on the chaperone machinery. We thus analyzed additional proteome parameters that represent expansion by innovation, rather than by mere duplication and divergence, as detailed below.
unusually high number of “aggregation hotspots” (defined as a poorly soluble protein segment of ≥5 aa in length, with solubility predicted from the sequence). The threshold for comparison was set at ≥20 hotspots per protein (at this threshold, ≤3% of proteins are aggregation-prone; the same trend was observed with lower thresholds; Dataset S8). We then calculated what percentage of the entire proteome these aggregation-prone proteins represent. This prediction is restricted by the fact that some of the predicted segments actually reside in the hydrophobic cores of stably folded proteins and hence do not comprise aggregation hotspots (37). However, such segments are likely to be as frequent in prokaryote and eukaryote proteomes (or even less frequent in the latter, where disordered proteins are abundant). Overall, our results indicate that compared with prokaryotes, aggregation-prone proteins have become nearly 6-fold more frequent in eukaryotes (Fig. 2E), with the highest frequency seen in Chordata proteomes (Fig. 2E).

Intrinsically disordered regions became ∼20-fold more frequent in the proteome. From prokaryotes to eukaryotes, did all the changes in proteome composition demand more chaperone action in the latter? The expansion of intrinsically disordered regions, that in principle would not demand increased chaperone action, could be an exception. Though explored before (38, 39), to have this expansion on the same scale and set of organisms used for analyzing all other proteome factors, each protein of the representative proteomes was scanned to infer disordered segments ≥100 aa long (Dataset S9). As plotted in Fig. 2F, from prokaryotes to eukaryotes, the percentage of proteome length that is disordered has expanded nearly 20-fold.

Overall, it appears that although gene and whole-genome duplications dominate, and in particular along with the evolution of eukaryotes (40), dramatic changes in proteome composition have occurred owing to bonafide innovations. Specifically, concerning the burden on the chaperone machinery, proteome compositions have changed massively with respect to new folds, beta superfolds, repeat proteins, fold combinations, and aggregation propensity.

The Evolutionary History of Chaperones. In parallel to estimating the expansion of proteome size and composition, we investigated the evolutionary history of chaperones, aiming to date their emergence and their expansion along the ToL. To that end, absence or presence, and copy numbers, of the core chaperones (HSP20, HSP60, HSP70, HSP100, and HSP90) was determined for the representative proteomes. Subsequently, protein trees were generated and compared with the ToL, to account for gene loss and horizontal transfer events. Protein sequences of all core-chaperone

Fig. 2. Expansion of proteomes by innovations. Figure features follow those of Fig. 1. (A) Shown on the y axis (log scale) is the average number of unique folds [ECOD X-groups (33)] in each phylogenetic clade. The lines were derived by a fit to an exponential equation, and are provided merely as visual guides. Prokaryotic (black and blue dots) and eukaryotic organisms (orange, gray, green, red, and wine dots) are largely separated by a dashed line in A. (B) Same as A, for the count of unique fold combinations. (C) Same as A, for the percentage of proteins in the proteome comprising at least one beta-superfold domain (note the linear scale). (D) Same as A, for the percentage of total proteome length that is repeated (also on a linear scale). (E) Same as A, for the percentage of proteins in the proteome predicted to have ≥20 aggregation hotspots per proteome (also on a linear scale; see also Dataset S10). (F) Same as A, for the percentage of proteome length that is intrinsically disordered (also on a linear scale).
families were extracted from the representative proteomes (Dataset S10). These sequences were aligned and used to generate maximum-likelihood, midpoint-rooted protein trees which were then compared with the ToL (SI Appendix, Table S1).

The core chaperones emerged in early-diverging prokaryotes. Our analysis traced the origin of all five core-chaperone families in early-diverging prokaryotes. The phylogenetic tree of a single protein, typically of a few hundred amino acids in length, often lacks the resolution required to reliably date the emergence, especially when horizontal transfer events are frequent. Dating emergence to LUCA is particularly challenging. We followed the recommendations of Berkemer and McGlynn (41) and demanded that for a chaperone family to be assigned to LUCA, there must be a single split between bacterial and archaeal sequences at the root of the protein tree, with strong bootstrap support for this split, and that interkingdom branches would be longer than the intrakingdom branches. These criteria assigned the emergence of only one core chaperone, HSP60—a cage-like ATP-fueled unfoldase (11, 42), to LUCA (Fig. 3 A). The protein tree of HSP60 further indicated an ancient horizontal gene transfer (HGT) from archaea to Firmicutes, as previously noted (43).

The protein tree of HSP20—an antiaggregation “holding” chaperone—depicted a clear single split of bacterial and archaeal domains at the root, albeit with weak bootstrap support, and interkingdom branch lengths were shorter than intrakingdom branch lengths (SI Appendix, Table S1). Similar uncertainties were noted for the majority of protein families formerly assigned to LUCA (41). Previous studies assigned HSP20 to LUCA (44, 45), which, despite the above uncertainties, we concur with (Fig. 3 A), although later emergence and HGT is an alternative. Indeed, in accordance with a previous study (46), the HSP20 protein tree suggests multiple HGT events, though given the weak bootstrap support it was difficult to distinguish between phylogenetic uncertainty and actual HGT events, let alone to assign donor and acceptor clades.

The remaining core-chaperone families, HSP70, HSP90, and HSP100, appear to have emerged in bacteria, though phylogenetic uncertainties, and probably extensive horizontal transfer between different bacterial clades and between bacteria and archaea, prevent the reliable assignment of their points of origin (Fig. 3 A). The ATP-dependent core chaperone HSP70—that controls protein unfolding, disaggregation, and degradation (47, 48)—was detected in the earliest-diverging bacterial clades Aquificae and Thermotogae.

---

**Fig. 3.** Evolutionary history of core and cochaperones. (A) The de novo emergence of core- and cochaperone families is summarized on the ToL. The ToL is the same as in Fig. 1A; clade names are omitted for clarity. Ancestral nodes in which a chaperone family emerged are marked with red stars and the core- and cochaperone families emerged in that node are listed. The dashed gray arrows reflect the endosymbiotic integration of archaeal and bacterial chaperone systems in LECA and eukaryotic and cyanobacterial chaperone systems in photosynthetic algae. (B) The percentage of core-chaperone genes per proteome (shown is the average percentage for each phylogenetic clade). Figure features follow Fig. 1. The line was derived by a fit to a linear equation and is provided merely as a visual guide. Prokaryotic (black and blue dots) and eukaryotic (light gray color) organisms are largely separated by a dashed line. (C) Core- and cochaperone gene expression, relative to ribosomal proteins, in cells. Plotted are model organisms for which sufficient, processed, and reliable expression data were available. The columns include core chaperones (colors represent the phylogenetic clades in Fig. 1A) and cochaperones (light gray color). The error bars represent the SD among different nonredundant abundance datasets. (D) Same as C, for the relative basal abundance of core and cochaperones in cells. (E) Same as B, for cochaperone genes per proteome. The line was derived by a fit to an exponential equation and is provided merely as a visual guide.
However, the protein tree clustered their HSP70 sequences with those from the later-diverging bacterial lineages Deltaproteobacteria, Clostridia, and Bacilli, with bootstrap values being too low to distinguish between phylogenetic uncertainty and true HGT events (SI Appendix, Table S1). Thus, HSP70 appears to have a bacterial origin, around or after the emergence of terrestrial bacteria, that is, around the divergence of Fusobacteria (49) (Fig. 3A). Following its emergence, HSP70 was likely horizontally transferred to archaea, but the current analysis could not reliably assign donor and acceptor clades.

The protein trees of both HSP90 and HSP100 depict a similar scenario. Both seem to have emerged in bacteria around or after terrestrial bacteria emerged (Fig. 3A). Although a reliable point of origin could not be assigned for either of these two chaperones, biochemical assays show that whereas the activity of HSP90 or HSP100 strictly depends on the presence of HSP70, HSP70 itself can act independently. Further, across the ToL, every organism that harbors genes for HSP90 and/or HSP100 also harbors genes for HSP70, but not vice versa. Thus, it is likely that HSP90 and HSP100 have both emerged after HSP70. Similar to HSP70, HSP90 and HSP100 were likely horizontally transferred to archaea. While our protein trees do indicate such trends, the bootstrap values are low.

The archael and bacterial core chaperones were integrated into the last euukaryotic common ancestor (LECA), and no new core chaperones emerged with the birth of eukaryotes (Fig. 3A). Chaperones of archael origin mostly continued to function in their original compartment, the cytosol. Although most Alphaproteobacterial endosymbiont genes were transferred to the nucleus, most of the chaperones of bacterial origin evolved to translocate back to the compartment from which they originated, namely to the mitochondria (50–53). Chaperone evolution in eukaryotes involved gene loss as well; for example, cytosolic and mitochondrial HSP100s have been lost in metazoans (53).

**The expansion of core chaperones.** Whereas no new core-chaperone family emerged in eukaryotes, gene copy numbers of the existing families did increase via gene duplication, to support expanding proteomes, for condition-specific expression, and also to cater for the emergence of multiple subcellular compartments. Bacteria and archaea typically harbor the same five core-chaperone families as eukaryotes. In any bacterial or archael genome, gene copy numbers of individual chaperone families range between 1 and 4, summing up to an average of 8 core-chaperone genes per proteome (Dataset S10). In comparison, the number of core-chaperone genes in higher plants, which are among the most complex eukaryotes, increased ~30-fold for HSP20, ~50-fold for HSP60, ~40-fold for HSP70, ~20-fold for HSP90, and ~10-fold for HSP100 (SI Appendix, Fig. S4). Parasitic microbes, such as Mycoplasma pneumoniae, Plasmodium falciparum, and Entamoeba histolytica, and photosynthetic bacteria, algae, and plants often harbor unusually high chaperone gene copy numbers, likely to counter the immune response of the host (54, 55) and the oxidative stress (56, 57). However, when proteome size is accounted for, it is evident that the expansion of core chaperones largely coincides with the overall expansion of proteome size. In fact, core chaperones comprise ~0.3%, that is, 3 out of 1,000 proteins, in all proteomes, from the simplest free-living prokaryotes to mammals (Fig. 3B). Further, the expansion of core chaperones occurred by gene duplication only, with no bona fide innovation, as all five core-chaperone families seem to have preexisted in prokaryotes.

**The cellular abundance of core chaperones increased ~6-fold.** As described above, the relative representation of core-chaperone genes is roughly the same in the genomes of prokaryotes and eukaryotes. However, gene expression levels could vary, and higher cellular levels of chaperones could support the increasingly complex eukaryotic proteomes. To assess expression levels, the messenger RNA (mRNA) and protein abundance of core or cochaperones was compared in prokaryotic and eukaryotic cells/tissues not subjected to stress or genetic modifications. Curated expression data could be obtained for 30 free-living model organisms spanning 14 major clades along the ToL (Dataset S11). The mRNA abundance of core-chaperone genes, relative to that of ribosomal proteins (as a proxy for the overall level of protein synthesis), has elevated ~6-fold in chordates compared with mesophilic bacteria (Fig. 3C). Protein abundance data available for 11 model organisms along the ToL (Dataset S12) indicate the same trend (Fig. 3D). Further, chaperone levels in eukaryotes are ~3-fold compared with extremophilic prokaryotes and pathogenic bacteria that in turn show ~2-fold higher levels than those of mesophilic nonpathogenic bacteria.

**Cochaperones expanded ~9-fold.** In eukaryotes, as the gene copy numbers of core chaperones expanded, and their protein abundance also increased, what happened to their auxiliary workforce, the cochaperones? The number of unique cochaperone families per proteome expanded from ~3 in prokaryotes to ~20 in humans. Most cochaperones are eukaryote-specific (Dataset S10), and therefore have likely emerged relatively recently, and only a few cochaperones are found in all three domains of life. To date their emergence, protein trees were generated. These suggest that, as expected, these cochaperones emerged after the core chaperone they work with. HSP60 is assigned to LUCA, while its bacterial cochaperone HSP10/GroES appears to have emerged later along with the emergence of bacteria (Fig. 3A). HSP70’s cochaperones, the J-domain proteins (JDPs) and GrpE, and HSP90’s cochaperone, Cyp40, appear to have emerged at the same node as their respective core chaperones. Given the phylogenetic uncertainties, possible extensive horizontal transfers, and the resolution that protein trees allow, a more precise dating could not be performed. It appears, however, that JDPS, GrpE, and Cyp40 have all emerged after the emergence of terrestrial bacteria (Fig. 3A), and therefore likely emerged after their respective core chaperones. Several cochaperone families emerged in eukaryotes (Fig. 3A), including HSP110 that diverged by duplication of HSP70 (58), and Pih1, Aha1, and Chip that harbor eukaryote-specific folds and hence likely emerged de novo. Overall, it is evident that cochaperones tail core chaperones, and not vice versa. With the birth of several cochaperones in eukaryotes, the percentage of genes encoding for cochaperones in the proteome expanded ~5-fold from prokaryotes to mammals (Fig. 3E). Notably, the JDPs, cochaperones of HSP70, are the major contributor to this copy-number expansion (17). Further, alongside the increase in copy numbers and the emergence of new families, the protein expression levels of cochaperones also increased across the ToL (cochaperone protein abundance is ~2-fold higher in chordates compared with mesophilic bacteria and mRNA abundance is ~4-fold higher; Fig. 3 C and D).

Our analysis, therefore, suggests that evolutionary innovation occurred primarily at the level of cochaperones that facilitated the basal core-chaperone activity, thus expanding the chaperone network to meet the challenges of newly emerging protein folds and increasingly complex proteomes.

**Discussion**

How did the proteomes expand across the ToL, and how did chaperones evolve to support this expansion? To address this question, we compiled data from multiple sources and analyzed them under one roof, thus allowing a systematic, quantitative comparison, as summarized in Fig. 4. From the simplest free-living prokaryotes to plants and animals, proteomes have continuously expanded by both duplications and innovations. It is primarily due to the latter that proteome “complexity” has continuously increased in various ways that demand increased chaperone action. Across the ToL and especially when comparing prokaryotes with eukaryotes, we see a larger number of proteins per proteome (Fig. 4A) as well as larger proteins. The latter relates to multidomain proteins being increasingly represented. The
number of different folds increases as well as of unique combinations of folds in multidomain proteins. Proteomes also contain a larger fraction of protein types that are prone to misfolding, such as repeat proteins and proteins comprising beta sheets, and those that are predicted to be highly aggregation-prone. The birth of a new fold, or of a new domain combination, likely results in poor foldability. With time, mutation and selection would improve foldability (59) and could ultimately render a newly born chaperone-independent protein. Nonetheless, the cumulative impact of newly evolved proteins, and of certain protein types (repeat or beta-rich proteins), likely demands increased chaperone capacity.

This dramatic increase in proteome complexity, and hence the demand for chaperone action, has not been met by the emergence of new core chaperones. Eukaryotes possess the same five core-chaperone families as prokaryotes, and metazoans and chordates have in fact lost HSP100 (Fig. 4B). Further, the relative representation of core-chaperone genes does not vary between prokaryotes and eukaryotes. Rather, the need for increased chaperone action was met in two ways: first, by an increased cellular abundance of core chaperones, and second, through the emergence of new cochaperones—while in bacteria ~4 cochaperone families are found, in eukaryotes their number increased to 15 or even 20 in Mammalia.

Most of the expanding proteome features are likely the outcome of adaptive evolution (e.g., the emergence of new folds and domain combinations). However, expansion may also occur by drift, that is, by fixation of genetic changes by chance, due to population bottlenecks. Indeed, the effective population size ($N_e$) has dropped from prokaryotes (typically >$10^8$) to unicellular
and substrate sharing between core chaperones are schematically completely processed HSP70 substrates (76, 77). Cooperativity can further be transferred to HSP60 for final refolding to the native substrates to HSP70 for ATP-driven unfolding, from which they can aggregated proteins and promote their subsequent refolding (72).

Incompletely processed misfolded or unfolded protein substrates, for example, HSP70 triggers a network that has emerged from a simple two-component system in this network (Fig. 4A). The main driving force for duplication and specialization is a functional trade-off—optimization of one function comes at the expense of other functions (67). However, given a “generalist” mode of function, the quality control of increasingly large and complex proteomes could be achieved by an elevated abundance of existing core chaperones, rather than by the emergence of new core-chaperone families. Indeed, although gene copy numbers of core chaperones have indeed increased by gene duplication, their relative representation compared with proteome size remained constant (Fig. 3B) and the resulting paralogous copies have mostly relocalized to different subcellular compartments or are expressed under different stress conditions (68). In parasitic microbes and photosynthetic organisms, paralogous copies have mostly relocalized to different subcellular compartments or are expressed under different stress conditions (68).

In eukaryotes, an integrated chaperone network, comprising both core and cochaperones, controls protein quality (69–71). In this network (Fig. 4C), the highly abundant core chaperones operate cooperatively, namely they not only share, and exchange incompletely processed misfolded or unfolded protein substrates, but also trigger the activities of another. HSP70 plays a critical role in this network by mediating cooperative communications between the other core chaperones. For example, HSP70 triggers the disaggregate activity of HSP100, and jointly they disaggregate aggregated proteins and promote their subsequent refolding (72–74). In another example, HSP20 can transfer misfolded substrates to HSP70 for ATP-driven unfolding, from which they can be further transferred to HSP60 for final refolding to the native state (75). Likewise, HSP90 can promote the maturation of incompletely processed HSP70 substrates (76, 77). Cooperativity and substrate sharing between core chaperones are schematically represented in Fig. 4C. Together, these generalist, cooperative core chaperones constitute the core of an integrated chaperone network that has emerged from a simple two-component system in LUCA (Fig. 4C).

Alongside the expansion of proteome complexity, the chaperone network has also expanded—primarily by the emergence of cochaperones (Fig. 4C). This expanding array of cochaperones augmented the ability of core chaperones to efficiently share substrates and to function cooperatively. In contrast to the generalist core chaperones, cochaperones are more diverse and accordingly seem to subspecialize in specific roles, including cochaperones that handle specific proteins. Examples include UNC45, a cochaperone that emerged in Fungi, and facilitates HSP90-mediated maintenance of myosin in metazoan skeletal and cardiac muscles (78). Another Fungi-born cochaperone, the Tsc1/2 heteromer, specializes in recruiting kinase and some nonkinase substrates to HSP90 (79). Other cochaperones mediate protein transport; examples include Tom70 and F23 that facilitate protein trafficking through Golgi and mitochondrial membranes (80–82). The specialist mode of function of cochaperones coincides with how they expanded, namely by duplication and divergence of ancient prokaryote-born cochaperones but also via bona fide innovations, namely by the emergence of completely new specialized cochaperones in eukaryotes. As shown here, the emergence of new cochaperones coincides with the emergence of new proteins (i.e., by de novo emergence rather than by duplication of preexisting proteins). However, co-occurrence does not mean coevolution—indeed, we know very little about the latter. Did certain cochaperones emerge to support the de novo emergence of a specific protein or protein class? If so, does chaperone dependency persist, hence making codependency a “selfish” irreversible trait? Alternatively, as some newly emerged proteins evolved further, their foldability improved, allowing them to become chaperone-independent.

Thus, across the Tree of Life, proteomes have massively expanded, not just by duplication of preexisting proteins but also by the emergence of completely new ones. Eukaryotic proteomes became particularly large and specifically richer in repeat, beta-rich, and aggregation-prone proteins whose folding is inherently challenging. These changes in proteome size and composition intensified the demand for chaperone action. Curiously, however, no new core chaperones emerged in response to this increased demand. Instead, they increased in abundance relative to all other proteins in the cell. Foremost, an entire network of cochaperones had evolved that facilitate the basal core-chaperone activity.

Materials and Methods

For details, see SI Appendix, Methods.

Proteome Size and Median Protein Length. A nonredundant set of 188 prokaryotic and eukaryotic organisms was collected from the TimeTree database (19) (listed in Dataset S1) and their proteome sequences (sequences of all proteins including splice variants, if relevant) were obtained from the National Center for Biotechnology Information genome database (83). For each organism, the total numbers of proteins in the respective proteome, and their lengths, were computed.

Multidomain Proteins. For each protein, domain annotations were collected from Pfam (32). The number of proteins comprising ≥3 and those comprising ≥3 domains were then counted per each proteome.

Number of Unique Folds and Fold Combinations. Proteins were clustered into independently evolved lineages using the ECOD database (which considers both sequence and structural similarities) where independently evolved lineages are termed X-groups (33). For each representative organism, the Pfam-assigned domains were mapped to their corresponding ECOD X-groups. The numbers of unique X-groups identified in a given proteome were considered as a measure of the number of unique folds. Similarly, we counted the X-group combinations, considering also their order along the polypeptide chain (AB ≠ BA) present in each protein. The total number of X-group combinations identified in a given proteome was considered as a measure of the total number of fold combinations.
Beta Superfolds in the Proteome. We counted how many proteins in the proteome contain at least one domain annotated as all-beta fold (annotated in EC numbers from the superfamily groups: beta barrel, beta meander, beta sandwich, beta duplicates or obligate multimers, and beta-complex topology). By normalizing this number by the total number of proteins in the proteome, we derived the fraction of beta-superfold proteins per proteome.

Repeated Sequences. Each protein in the representative proteomes was scanned by the T-REKS repeat-identifier program (84) to detect repeats that are ≥ 50 aa in length, with solubility predicted from the protein’s sequence using CasSol v2.1 (85). For each representative proteome, we computed the percentage of proteins in the proteome that contain ≥ 20 aggregation hotspots.

Aggregation-Protein Probes. An aggregation hotspot was defined as a “poorly soluble” protein segment of ≥ 50 aa in length, with solubility predicted from the protein’s sequence using CasSol v2.1 (85). For each representative proteome, we computed the percentage of proteins in the proteome that contain ≥ 20 aggregation hotspots.

Intrinsically Disordered Regions. Intrinsically disordered regions were identified by scanning all proteins in the representative proteomes by lPred2A (86). Disordered segments ≥ 100 residues were considered. The percentage of proteome length that is disordered was subsequently derived (the total number of residues assigned to disordered segments multiplied by 100, divided by the sum of the length of all proteins).

Evolutionary History of Chaperones. To determine the evolutionary appearance and expansion of the core-chaperone and cochaperone families, we identified their occurrences in the 188 representative organisms, using two complementary methods. The first method involved manual curation of annotated chaperones in model organisms that were subsequently used as queries to find orthologous sequences in the other organisms by protein-protein BLAST (87). The second method involved identifying the Pfam-assigned domain combinations of the various known chaperones in model organisms. Subsequently, any protein in the representative proteomes comprising these domain combinations was assigned as a member of the corresponding chaperone family. The orthologous and paralogous sequence families were assigned using MUSCLE v3.8.31 (88). Maximum-likelihood phylogenetic trees were generated using MEGA X (89). To date the emergence of individual chaperone families, the protein trees were manually compared with the ToL to assign the node of emergence and possible HGT events.

Chaperone Abundance Analysis. To quantify the variation in chaperone mRNA abundance, RNA sequencing (RNA-seq)-based expression data for various model organisms were collected from available resources. Only processed RNA-seq data were considered where the normalized abundances of mRNA transcripts are provided as transcripts per million (TPMs). For each experiment, the sum of the TPM values of the core and the cochaperones was divided by the sum of TPMs of all ribosomal proteins. The average and SD over all experiments per given organism were computed.

To quantify protein abundance, mass spectrometry-based protein abundance data were collected for various model organisms from PaxDb (90). For each dataset, the sum of abundance values of all core chaperones and cochaperones was normalized by the sum of abundance values of all other proteins. The average and SD over all experiments per given organism were computed.

Data Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Jagoda Jablonska for helping to establish the representative Tol, and Iia Gruic-Sovulj for proposing the chaperone abundance analysis. We thank Rina Rosenzweig, Harm H. Kampinga, Matthias P. Mayer, Sudip Kundu, Marc Robinson-Rechavi, and Agnieszka Klosowska for valuable suggestions regarding the manuscript. This work was supported by a Minerva Foundation grant (to D.S.T.) and Grant 31003A, 175453 from the Swiss National Fund (to P.G. and M.E.R.). S.M. was supported by a postdoctoral fellowship, provided by Israel’s Council for Higher Education Planning and Budgeting Committee. P.G. is an associate professor in the Department of Plant Molecular Biology at the University of Lausanne. D.S.T. is the Nella and Leon Benoziyo Professor of Biochemistry at The Weizmann Institute of Science.

1. C. Woese, The universal ancestor. Proc. Natl. Acad. Sci. U.S.A. 95, 6854–6859 (1998).
2. W. F. Doolittle, Phylogenetic classification and the universal tree. Science 241, 2124–2129 (1989).
3. L. Eme, A. Spang, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. Nat. Rev. Microbiol. 15, 711–722 (2017).
4. G. M. Cooper, The Cell: A Molecular Approach (Sinauer Associates, 2000).
5. A. Finka, R. U. Matteo, P. Goloubinoff, Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. Annu. Rev. Biochem. 74, 686–688 (2008).
6. N. A. Heim et al., Hierarchical complexity and the size limits of life. Proc. Natl. Acad. Sci. U.S.A. 20171039 (2017).
7. C. B. Anfinsen, Principles that govern the fold of protein chains. Science 181, 223–230 (1973).
8. K. A. Dill, H. S. Chan, From Levinthal to pathways to funnels. Nat. Struct. Biol. 4, 10–19 (1997).
9. S. W. Englander, L. Mayne, The nature of protein folding pathways. Proc. Natl. Acad. Sci. U.S.A. 111, 15873–15880 (2014).
10. J. H. Han, S. Batty, A. A. Nickson, S. A. Teichmann, J. Clarke, The evolution and folding of multidomain proteins. Nat. Rev. Mol. Cell Biol. 8, 319–330 (2007).
11. A. Finka, R. U. Matteo, P. Goloubinoff, Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. Annu. Rev. Biochem. 80, 323–355 (2013).
12. A. J. Caplan, What is a chaperone? Cell Stress Chaperones 8, 105–107 (2003).
13. F. H. Schopf, M. M. Biebl, J. Buchner, The HSP90 chaperone machinery. Nat. Rev. Mol. Cell Biol. 18, 345–360 (2017).
14. E. J. Duncan, M. E. Cheetham, J. P. Chappell, J. van der Spey, The role of HSP90 and its co-chaperones in protein misfolding, aggregation and disease. Subcell. Biochem. 78, 243–273 (2015).
15. N. Tokuriki, D. S. Tawfik, Protein dynamism and evolvability. Science 324, 203–207 (2009).
16. D. Bogumil, T. Dagan, Cumulative impact of chaperone-mediated folding on genome evolution. Biochemistry 51, 9941–9953 (2012).
17. D. Alvarez-Ponce, J. Aguilar-Rodriguez, M. A. Fares, Molecular chaperones accelerate the evolution of their protein clients in yeast. Genome Biol. Evol. 11, 2360–2375 (2019).
18. S. B. Hedges, J. Marin, M. Sulecki, M. Paymer, S. Kumar, Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. 32, 835–845 (2015).
19. C. Wang, V. N. Uversky, L. Kurgan, Disordered nucleotides: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16, 1486–1498 (2016).
20. P. Mayer, Sudip Kundu, Marc Robinson-Rechavi, and Agnieszka Klosowska for valuable suggestions regarding the manuscript. This work was supported by a Minerva Foundation grant (to D.S.T.) and Grant 31003A, 175453 from the Swiss National Fund (to P.G. and M.E.R.). S.M. was supported by a postdoctoral fellowship, provided by Israel’s Council for Higher Education Planning and Budgeting Committee. P.G. is an associate professor in the Department of Plant Molecular Biology at the University of Lausanne. D.S.T. is the Nella and Leon Benoziyo Professor of Biochemistry at The Weizmann Institute of Science.

Rebeaud et al. On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life

11. A. Finka, R. U. Matteo, P. Goloubinoff, Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. Annu. Rev. Biochem. 80, 323–355 (2013).
12. A. Finka, R. U. Matteo, P. Goloubinoff, Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. Annu. Rev. Biochem. 80, 323–355 (2013).
13. A. J. Caplan, What is a chaperone? Cell Stress Chaperones 8, 105–107 (2003).
14. F. H. Schopf, M. M. Biebl, J. Buchner, The HSP90 chaperone machinery. Nat. Rev. Mol. Cell Biol. 18, 345–360 (2017).
15. E. J. Duncan, M. E. Cheetham, J. P. Chappell, J. van der Spey, The role of HSP90 and its co-chaperones in protein misfolding, aggregation and disease. Subcell. Biochem. 78, 243–273 (2015).
16. N. Tokuriki, D. S. Tawfik, Protein dynamism and evolvability. Science 324, 203–207 (2009).
17. D. Bogumil, T. Dagan, Cumulative impact of chaperone-mediated folding on genome evolution. Biochemistry 51, 9941–9953 (2012).
18. D. Alvarez-Ponce, J. Aguilar-Rodriguez, M. A. Fares, Molecular chaperones accelerate the evolution of their protein clients in yeast. Genome Biol. Evol. 11, 2360–2375 (2019).
19. S. B. Hedges, J. Marin, M. Sulecki, M. Paymer, S. Kumar, Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. 32, 835–845 (2015).
20. L. Margulis, M. Chapman, R. Guerrero, J. Hall, The last eukaryotic common ancestor (LECA): Acquisition of cytoskeletal motility from aerotolerant sprochetes in the Proterozoic Eon. Proc. Natl. Acad. Sci. U.S.A. 103, 13080–13085 (2006).
B. Xue, A. K. Dunker, V. N. Uversky, Orderly order in protein intrinsic disorder distribution: Disorder in 35000 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. 30, 137–149 (2012).

Y. Van de Peer, S. Maere, A. Meyer, The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. 10, 725–732 (2009).

J. S. Berkemer, S. E. McFall, A new area of Archaea-Bacteria domain separation: Variable phylogenetic distance and the tempo of early evolution. Mol. Biol. Evol. 37, 2323–2340 (2020).

H. Salehi, Chaperone machines for protein folding, unfolding and disaggregation. Nat. Rev. Mol. Cell Biol. 14, 630–642 (2013).

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

F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. Biochim. Biophys. Acta 1857, 1027–1038 (2016).

M. C. Weiss et al., The physiology and habitat of the last universal common ancestor. Nat. Microbiol. 1, 16116 (2016).

T. Kriehuber et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

F. U. Battistuzzi, S. B. Hedges, A major clade of prokaryotes with ancient adaptations due to indels in intrinsically disordered regions. Mol. Biol. Evol. 33, 531–543 (2016).

M. R. Woodford et al., Guidelines for the nomenclature of the human heat shock proteins. Cell Stress Chaperones 14, 105–111 (2009).

J. U. Dahl, M. J. Gray, U. Jakob, Protein quality control under oxidative stress conditions. Cold Spring Harb. Perspect. Biol. 12, a033951 (2020).

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

S. M. Hemmingsen et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

R. S. Altschul et al., Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

T. Kriehuber et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

R. S. Altschul et al., Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

R. K. Fisher, R. R. Rollins, The interaction of the Hsp70 chaperone determined by screening cellulose-bound peptide libraries. EMBO J. 16, 1501–1507 (1997).

D. S. Tawfik, L. Gruic-Sovulj, How evolution shapes enzyme selectivity—Lessons from aminoacyl-tRNA synthetases and other amino acid utilizing enzymes. FEBS J. 287, 1284–1305 (2020).

H. H. Kampinga et al., The physiology and habitat of the last universal common ancestor. Nat. Microbiol. 1, 16116 (2016).

S. J. Berkemer, S. E. McFall, A new area of Archaea-Bacteria domain separation: Variable phylogenetic distance and the tempo of early evolution. Mol. Biol. Evol. 37, 2323–2340 (2020).

H. Salehi, Chaperone machines for protein folding, unfolding and disaggregation. Nat. Rev. Mol. Cell Biol. 14, 630–642 (2013).

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

F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. Biochim. Biophys. Acta 1857, 1027–1038 (2016).

M. C. Weiss et al., The physiology and habitat of the last universal common ancestor. Nat. Microbiol. 1, 16116 (2016).

T. Kriehuber et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

F. U. Battistuzzi, S. B. Hedges, A major clade of prokaryotes with ancient adaptations due to indels in intrinsically disordered regions. Mol. Biol. Evol. 33, 331–334 (2009).

B. Bukau, A. L. Horwich, The Hsp70 and Hsp60 chaperone machines. EMBO J. 33, 1497–1502 (2014).

R. S. Altschul et al., Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

T. Kriehuber et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

R. S. Altschul et al., Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

T. Kriehuber et al., Independent evolution of the core domain and its flanking sequences in small heat shock proteins. J. Mol. Biol. 433, 682 (2018).

R. S. Altschul et al., Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).