Protein homeostasis imprinting across evolution

Thodoris Koutsandreas[1,2], Brice Felden[3,6,7], Eric Chevet[4,5,6*] and Aristotelis Chatziioannou[1,2,6*]

[1]Center of Systems Biology, Biomedical Research Foundation of the Academy of Athens, Athens, Greece. [2]e-NIOS Applications PC, Kallithea-Athens, Greece. [3]University of Rennes, INSERM U1230, Rennes, France. [4]INSERM U1242, University of Rennes and [5]Centre de Lutte Contre le Cancer Eugène Marquis, Rennes, France.

Keywords: Protein homeostasis, Molecular evolution, Bacteria, Archaea, Eukarya, Network-driven phylogenetic analysis

[6]equally contributed to this work

[7]this article is dedicated to the memory of Pr Brice Felden

****Corresponding Authors:*** EC (eric.chevet@inserm.fr), AC (achatzi@eie.gr)

Abstract

Protein homeostasis (a.k.a. proteostasis) is associated with the primary functions of life, and therefore with evolution. However, it is unclear how the cellular proteostasis machines have evolved to adjust the protein biogenesis needs to environmental constraints. Herein, we describe a novel computational approach, based on semantic network analysis, to evaluate proteostasis during evolution. We show that the molecular components of the proteostasis network (PN) are reliable metrics to deconvolute the life forms into Archaea, Bacteria and Eukaryotes and to assess the evolution rates among species. Semantic graphs were used as a new criterion to evaluate PN complexity of 94 Eukaryotes, 250 Bacteria and 93 Archaea, thus representing a novel strategy for taxonomic classification. This functional analysis provided information about species divergence and pointed towards taxonomic clades that evolved faster than others. Kingdom-specific PN were identified, suggesting that PN complexity correlates evolution. Through the analysis of gene conservation, we found that the gains or losses that occurred throughout PN evolution revealed a dichotomy within both the PN conserved modules and within species-specific modules. Since the PN is implicated in cell fitness, aging and disease onset, it could be used as a new metric to tackle mechanisms underlying 'gain-of-functions', and their biological ramifications.

Introduction

Protein homeostasis (a.k.a. proteostasis) refers to a complex and interconnected network of processes that affects both expression levels and conformational stability of proteins in cells by controlling their biogenesis, folding, trafficking and degradation within and outside the cell. The molecular mechanisms controlling proteostasis, are implicated in cell fitness, aging and contribute to disease onset. The underlying network of cellular mechanisms (i.e. the proteostasis network - PN) includes protein synthesis, co/post-translational folding, quality control, degradation, as well as adaptive signaling in response to proteostasis imbalance (Balch et al., 2008). From Prokaryotes to Eukaryotes, the PN was subjected to evolutionary pressure for each organism, to cope with intrinsic and extrinsic demands. Evolution was shaped by factors such as genome complexity, post-translational modifications repertoire, the presence of subcellular compartments, the emergence of multicellular organisms and cell differentiation with cells exhibiting high protein synthesis yields and secretion requiring a robust PN. An effective PN is also instrumental for eukaryotic cells with temporal variations of protein expression (e.g. neurons and endocrine cells). Each of these constraints increased the needs for updated, adaptive mechanisms, ensuring protein homeostasis (Roth & Balch, 2011). As such, the PN was subjected to selections specific to each organism. For instance, compartment-specific proteostasis control machineries in the cytosol, the endoplasmic reticulum (ER) or the mitochondria are required in eukaryotic cells, whereas dedicated systems are found in the periplasm of Gram-negative Bacteria (Powers & Balch, 2013).

Traditional phylogenetic approaches use the sequence of conserved genes or proteins (or groups of them) as standard references rather than their functional identities to form clades, ancestral lineages and identify speciation events. As such, the construction of PN-related phylogenetic dendrogram requires to take into consideration the diversity of PN across species of various taxa. In theory, heat shock proteins (HSPs), that exert fundamental roles in maintaining protein homeostasis, could be considered as the appropriate markers to reflect PN evolution. Some HSP families (e.g. HSP40 and HSP70) are highly represented in most cells and the nature of this representation might reflect the underlying evolutionary relationships - e.g. 3 members of HSP40 in *E. coli* and 49 members in *Homo sapiens* (Gur et al., 2005; Kampinga et al., 2009). However, the informational content of their nucleotide and amino acid sequences remains unable to provide insights about functional evolution of proteostasis, hence the need for a different vocabulary to exploit their functional profiles and the subsequent structured networks.

Even if studies have reported computational models of proteostasis in *E. coli* (Powers et al., 2012) and in Eukaryotes (Wiseman et al., 2007), an overall layout of proteostasis evolution is still

lacking. Herein, we propose a novel method for phylogenetic analysis based on the semantic network analysis of ontological graphs, which describes the functional characteristics of species, under a specified cellular process, such as proteostasis. Using this approach to evaluate the functions of specific biological modules (i.e. list of genes/proteins), we assessed how proteostasis evolved with the phylogeny. Proteostasis-based evolutionary maps performed well *vs.* the gold-standard ribosomal RNA (rRNA) sequence-based evolution metric (Gilbert, 1986; Smit et al., 2007). Specifically, these maps managed to yield a phylogenetic dendrogram, which proposes taxonomic classifications of organisms based on the topology of their PN. In addition those newly generated dendrograms generally reproduce the established classifications in major taxa. Species allocation into distinct clusters populated by members from various phyla unveiled new shared adaptive responses to environmental cues. We show here that proteostasis, as a whole, is a reliable metric for species partitioning, providing a snapshot of the overall functional diversification of cellular functionality in the tree of life. Conserved and 'species-specific' PN components were identified, implying that the emergence of cell compartments was a major driving force in proteostasis expansion and diversification.

Results

Process implemented to evaluate evolution of the proteostasis network - As proteostasis maintenance is essential to warrant proper proteome functions, it represents *de facto* a conserved actor during evolution. We sought to study how the PN evolves among organisms, taking into account functional differences and commonalities recorded into the topology of its different architectures. This could provide a scaffold enabling the delineation of a PN-derived phylogenetic tree. First, we built the PN semantic profile for thousands of species, based on the Gene Ontology Biological Process (GO-BP) annotation of gene lists related to proteostasis. We also used the available rRNA nucleotide sequences and heat shock protein (HSP) amino acid sequences, to perform phylogenetic analyses on the same group of organisms, allowing therefore their comparison with PN-based trees. An exhaustive analysis of biological data public repositories (Ensembl (Cunningham et al., 2018), UniprotKB (The UniProt Consortium, 2018), and ENA (Leinonen et al., 2010)) was performed to collect the raw data. Organisms were included in the analysis by meeting the following criteria: i) adequate genomic annotation in the GO-BP corpus; ii) availability of gene sequences of 16S (for Prokaryotes) and 18S (for Eukaryotes) rRNAs and iii) at least one annotated amino acid sequence of HSP40 (*dnaJ*) and HSP70 (*dnaK*). Using both manual and automated procedures (see Materials & Methods), we generated comprehensive lists of rRNA, HSP40, HSP70 sequences and a gene list related to the PN for 437 organisms (94 Eukaryotes, 250 Bacteria and 93 Archaea; **Figure 1A, step II**). Pathway analysis was used to translate each gene list into GO-BP terms, which imprinted the PN-associated network for each species (**Figure 1B, step I,** for gene lists and pathway analysis results see **Code and Data Availability** section). A phylogenetic comparison of these functional profiles was performed through the calculation of the semantic PN similarities. Specifically, we used a standardized version of GO-BP in conjunction with semantic operators, to quantify the similarities of the term lists and create a phylogenetic dendrogram by clustering together species with high semantic similarities (**Figure 1B, step II**). Such dendrogram entailed the use of a reference version of GO-BP, for the calculation of the inter-species topological similarities, thus neutralizing annotation imbalances which incurred bias in the description of proteostasis among different organisms. This was achieved by the exclusion of specific terms, residing low in densely populated ontological tree branches, through the application of cutoff criteria for two metrics, Information Content (IC) and Semantic Value (SV). Using such an approach, a graph of 2000 terms was generated (**Figure 1C**). Finally, the PN-based phylogenetic tree, as well as those of rRNA and HSP sequences, were built using agglomerative clustering with the Ward's minimum variance method (Ward, 1963).

Ribosomal RNA, HSP40, HSP70 and PN-based phylogenies - Phylogenetic trees (**Figures S1-4**) were constructed to compare rRNA nucleotide sequences, HSP40, HSP70 amino-acid sequences and the graph topology of PN conservation. To further explore the characteristics of each evolutionary tree, organisms were projected on a two-dimensional space using the Multidimensional Scaling (MDS) algorithm (Borg &Groenen, 2005). Both rRNA sequences and PN functional profiles produced independent sub-groups for each established taxonomic domain (**Figure 2A**). The sole inconsistencies with the reference taxonomy from rRNA-based phylogenetic tree were observed with a group of *Streptophyta,* which clustered near Bacterial species, and with PN-based classification that showed five Archaea (*Methanobrevibacter* and *Methanothrix genera* and the other ones from ammonia-oxidizing species) embedded into the bacterial cluster (**Figure 2A, Figures S1-4**). PN evolution appeared less constrained than that of rRNA, which led to distantly separated super-kingdoms. This probably reflects the heterogeneity of the PN content, bisected into domain-specific components and others that are conserved across evolution. Nevertheless, the pairwise distances among species for rRNA and PN showed unambiguous correlation (**Figure 2B, Figures S5-6**). Concerning the HSP-derived classification, a poor correlation of the HSP40 and HSP70 sequences with evolution was observed, as they only succeeded to separate eukaryotic and prokaryotic kingdoms, even so not flawlessly. A weak but not negligible correlation with rRNA sequence-based distances suggests that taxonomic clusters intra-distances follow approximately the same distribution to their inter-distances (**Figures S7-8**). These observations imply that HSPs, which are individual components of the PN, lack informative power as marker of species evolution.

We next sought to compare PN's ability to classify species of the same taxonomic domain by comparison to what was achieved using rRNA and HSPs sequences. As such, we studied 'class-level' categorization of Archaea and Phylum groups of Eukaryotes and Bacteria. For each criterion, different grouping models of species, for a range of predefined number of clusters were generated, and the consistency of each model with the reference classification was assessed, using the Homogeneity Score (see Materials & Methods). The output corroborated the findings inferred from the rRNA sequences and revealed an overall homogeneity of PN profiles in Bacteria and Archaea (**Figure S9**). None of the criteria succeeded to produce an optimal number of clusters equal to that of the reference taxonomic groups, hence verifying that species of different lower-level taxonomies share highly similar profiles. In conclusion, these results show that rRNA sequences represented the most effective measure and that Bacteria and Archaea shared PN components. In addition, the eukaryotic PN was more complex, encompassing variations that coherently segregated the species, rescuing phyla segregations. HSP-derived clusters were similar to those obtained with rRNA

sequences in Prokaryotes, but declined among Eukaryotes, probably due to the high variation of protein families possessing 'species-specific' profiles. At last, PN-derived trees provided a much more informative and robust classification at the Phylum level in Eukaryotes, compared to either HSP40- or HSP70-derived trees (**Figure S9**).

Proteostasis as a functional metric to trace evolution - Pathway analysis using the GO-BP annotation and the subsequent semantic clustering of the enriched terms indicated that the major biological processes could be classified into 56 distinct groups. Terms associated with protein metabolic process and folding were represented in all tested species, thereby representing the PN "conserved core" (**Figure 3A**). In addition, specific processes related to programmed cell death and signaling pathways were enriched in Eukaryotes. Enrichment in the response to endoplasmic reticulum (ER) homeostasis imbalance in Eukaryotes was effective in all the tested species, as expected pointing out this compartment as a hotspot for proteostasis. This was evidenced by functions of the PN associated with protein glycosylation or carbohydrate metabolism. Similarly, enrichment of membranous organelles-associated mechanisms, such as mitochondrial organization or autophagy, were exclusively identified in Eukaryotes, whereas metabolic pathways such as nitrogen compounds processes, amide metabolism or even protein unfolding were associated to Prokaryotes (**Figure 3A**). We next constructed a PN conservation tree in Eukaryotes, Bacteria and Archaea to identify the commonalities and discrepancies between these kingdoms (**Figure 3B**). We used the association of species with each GO-BP cluster, to derive their two-dimensional representations for the whole PN (**Figure 3C**, left), the common PN (**Figure 3C**, middle) and the differential PN (**Figure 3C**, right). PN profiling produced independent sub-groups for each taxonomic domain. As predicted, these sub-groups increased their density when analyzing the PN "conserved core" (**Figure 3C**, middle) and seemed sparser when analyzing the differential PN (**Figure 3C**, right). Moreover, analysis of the differential PN unveiled a great diversity within each domain. In the case of Archaea, where proteostasis was determined by the "conserved core" PN, the great diversity might were caused by nuances in the differential components or potential false positives during the pathway analysis, which subsequently induced great semantic distances. Collectively, these data unveil conservation and divergence within the PN, associated with groups of ubiquitous functions and functions specific to each taxonomic domain.

Impacts of PN evolution on other functional networks - To evaluate the robustness of our methodology and to analyze the impact of the PN on the evolution of other cellular pathways, we

examined other conserved mechanisms, using the same procedure, as that described above. The ability of each conserved mechanism to separate the main taxonomic domains was evaluated using the Homogeneity (HS) and Silhouette (SS) scores (see Materials & Methods). To this end, we analyzed 20 conserved cellular pathways, as defined in the GO-BP, and both semantic analysis and phylogenetic tree construction were performed based on their whole pathway annotation. Then PN components, which were comprised in those pathways, were excluded to repeat this analytic process upon their exclusion. This showed that biological processes, which emerge as gains of functions during species evolution, or those, which evolved through higher functional complexity, segregate well the three super-kingdoms (**Figure 4; S10A-S17A**). The biological processes identified through this approach are, for instance, tRNA processing, cell compartmentalization, regulatory networks, lipid metabolism and methylation. Some performed marginally better in terms of accuracy, as taxonomic metric, compared to proteostasis or to the rRNA sequences - i.e. slightly higher HS values due to the erroneously classified Archaea in the case of PN, and the plants group in the case of rRNA sequences. Nevertheless, the phylogenies of these biological processes exhibited lower SS values than those obtained for PN (**Figure 4B**), which means that the produced clusters were sparser (especially within Prokaryotes) and consequently, the phylogenetic trees contained broader clades. In addition, parts of the aforementioned processes or those with narrower functional networks showed weaker performance, concerning clustering efficiency, especially due to their strong commonalities among the Prokaryotes. HS measurements were below 0.8, because many Archaea sequences were classified in Bacteria, and *vice versa* (**Figure 4A; S18A-S29A**). To evaluate the contribution of the PN to those machineries, we artificially removed proteostasis-related components from each process, which affected the results (**Figure 4; S10B-29B**). Only the regulatory processes (regulation of cellular amide metabolic process, regulation of protein metabolic process, post-transcriptional regulations of gene expression) and cellular component assembly retained adequate information to distinguish accurately the taxonomic domains, implying that their mechanistic frameworks can differentiate the taxonomic kingdoms. In general, all the processes with accurate performance suffered from low Silhouette scores, and some lost their phylogenetic congruity. Improvements in HS and SS values for some biological processes in the defective group were meaningless, as they did not address the phylogenetic comparison. This analysis indicates that the PN contributes to interconnected pathways involved in cell homeostasis and functions.

Discussion

Taxonomies based on molecular sequences have increased our understanding of evolutionary processes. Phylogenies based on isolated macromolecules conserved among evolution, such as nucleic (DNA/RNA) or protein sequences, have their limits as they do not accurately represent the complexity of life evolution. In this study, we designed and applied a novel approach for phylogenetic comparison which uses the semantic graph as new metric, to evaluate the complexity of protein homeostasis through monitoring the proteostasis network (PN). To the best of our knowledge, it the first time that semantic network analysis is used to define an evolutionary marker. This stands as a novel strategy for taxonomic classification, which assess the divergence of the semantic topology, providing species-specific PN profiles rather than analyzing individual sequences. The crux of this new method is the semantic comparison of the inferred ontological graphs. Using a standardized ontological framework, the semantic interpretation of gene/protein sets for different species provided biological insights about the impacts of the evolutionary pressure, and the extent of conservation of mechanisms among Eukaryotes, Archaea and Bacteria. In this premise, our study aimed at illustrating the complexity of the PN modular architecture. In addition, our analysis measured species-specific topological differences, and translated those into evolutionary paths according to the adaptive constraints. We show that monitoring PN characteristics provides reliable information about the evolution of living organisms, whereas monitoring its individual components has limited interpretation.

First, PN succeeded to separate the three main taxonomic domains almost infallibly, performing as accurately as rRNA sequences do. Moreover, PN performed much better than isolated PN components, such as the heat shock proteins (**Figure 2**). This primarily indicates that the proposed new method is able to disclose evolutionary differences among species of different taxa. It also demonstrates the utility of PN as a reliable evolutionary marker, able to classify species according to their main taxonomy, contrary to the limitations of the sequence-based approaches. The efficiency of the PN metric to increase taxonomic resolution dropped in Bacteria and Archaea, likely due to their poor annotations compared to the Eukaryotes. The evolution plasticity among Prokaryotes highlights the need for systematic annotations through rigorous measures, for instance with PN topology, that could provide objective rather than subjective classifications. GO-BP provides a descriptive genomic annotation which might is not suitable to elucidate slight functional differentiations between species. However, the impediment is that there is not any controlled vocabulary which describes the universe of cellular functionality in the level of pathways' topology for a plethora of species. For instance, the BioCyc database (Karp et al., 2019) contains curated and

computationally inferred annotation for thousands of species only for metabolic pathways, while Reactome Pathways (Jassal et al., 2020) provides curated annotation for the network of pathways of a few model-species. A way to solve this issue could be the massive integration of data from different vocabularies into a unified ontological schema. In addition, the weaker segregating capacity of the PN in the two Prokaryotic kingdoms might also indicates that their evolutionary pressure was managed through the functional diversification of selected genes (e.g. mutagenesis) or chromosome variations. In Eukaryotes, the complexity and adaptability of molecular circuitries became a major driving force. For the eukaryotes, the evolution of compartmentalization impinges upon the complexity of protein circuitry, prompting evolution of the PN to cope with these additional constraints (Powers & Balch, 2013).

Deconvolution of the PN in various molecular sets revealed that many, if not all cellular processes, are connected to the PN, especially for Eukaryotes. The conserved PN component across the vast majority of species is linked with protein production, folding and degradation, as well as responses to external or internal stimuli, activation or repression of anabolic and catabolic processes, to maintain cell homeostasis, as well as the proper localization of macromolecules. This comprises HSPs, which perform poorly in terms of separating the taxonomic kingdoms, but were recently shown to organize beyond the "de novo versus stress-inducible" scheme into a layered core-variable architecture in multi-cellular organisms (Shemesh et al., 2021). In addition to this, in order to cope with proteome complexity that arose with evolution, conserved core chaperones increased in abundance and new co-chaperone families appeared (Rebeaud et al., 2021). Other functional modules, considered as 'gain- or loss-' of adaptive functions, were also found to be connected to the PN. This, in turn, could help identify the role of PN in maintaining those functional changes (**Figure 4**). The instrumental role of proteostasis as a robust indicator of cellular and organismal adaptation to evolutionary cues, is highlighted by the taxonomic underperformance that the other mechanisms linked to PN exhibit when PN components are excluded (**Figure 4**). This provides evidence for a tight coupling between proteostasis and other major biological processes, and to consider PN as a trigger for functional diversification, particularly in Eukaryotes. As such, PN encompasses all the biological information to categorize species according to their complexity and acts as a clear-cut fingerprint of evolution, as it has co-evolved with the cell proteome and provided a driving force for adaptation to favor emergence of new traits.

As we show here that PN is a novel and reliable evolution metric, we also thought about the use of this integrated information as a quantitative trait to categorize diseases, and possibly their treatments. Approaches relying on the analysis of PN sub-networks (e.g. the "chaperome") were

proposed to be effective in various pathologies, such as cancers or degenerative diseases (Brehme et al., 2014; Hadizadeh Esfahani et al., 2018; Taldone et al., 2014), or in cell differentiation (Vonk et al., 2020). Considering that the "chaperome" would represent an evolutionary conserved part of the PN, one might envision an approach relying on the PN, as defined in our study, to assess how its deregulation could mark disease appearance, evolution, or even unveil new therapeutic targets. At last, at a time of an unprecedented world pandemic with SARS-CoV2, one might also consider PN evolution, and its quantitative measurement, as a tool to predict the evolution of human and animal pathogens. The evolution of influenza virus is affected by the targeted alteration of the host's cells PN, mostly through perturbations of HSF1 and HSP90 expression, once again belonging to the conserved core of the PN (Phillips et al., 2017). It may illustrate the needs for any pathogens to rely on a given host's PN, which could be evaluated using molecular data sets (e.g. protein-protein interactions, gene expression). Our results could pave the way to investigate PN alterations and their consequences for the physiology and pathology of their hosts.

Materials & Methods

Proteostasis-related genes lists and data acquisition – The PN encompasses various mechanisms, pertinent to different functional aspects, as protein quality control, production, concentration maintenance and degradation. Moreover, as proteins constitute essential functional cellular entities, proteostasis intricately regulates a multitude of cellular processes and consequently has a powerful contribution to the large phenotypic diversity observed. This large complexity is only partly recorded by various, available vocabularies, describing biological pathways and processes (e.g. Gene Ontology (The Gene Ontology Consortium, 2019), Reactome (Jassal et al., 2020), KEGG (Kanehisa & Goto, 2000)). In this sense, the semantic representation and annotation of the PN largely skewed, either in terms of functional vocabulary or at the organism level. On account of this, we defined species-specific gene lists, related to proteostasis, in order to reveal the PN components utilizing the pathway analysis. This gene selection procedure was performed through a supervised multi-phase analytic workflow (**Fig 1A, step I**), and it aimed to include genes strongly associated with key functional components, such as protein folding, degradation, endoplasmic reticulum, autophagy and associated signaling pathways. In this way, seed gene lists were inferred for eight eukaryotic model organisms (*Homo sapiens, Gallus gallus, Danio rerio, Xenopus tropicalis, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae and Arabidopsis Thaliana*) and a generic gene list was issued for the prokaryotic domain. This provided the initial scaffold, allowing to increase the number of organisms. For Eukaryotes, we applied homology mappings so as to retrieve putative, functionally similar genes, from the Ensembl repositories (vertebrates, fungi, metazoans and plants, (Kinsella et al., 2011)). The aforementioned model organisms were used to detect homologies with species belonging to the same taxonomic classification level (e.g. S. *cerevisiae* was used as reference organism for fungi and *A. thaliana* for plants). Hence, representative proteostasis-related gene lists were constructed automatically for hundreds of Eukaryotes. However, in order to exclude spurious annotations, only species pertaining genes with more than 50% homology to the reference gene lists were included into the final set.

For Prokaryotes, an advantage of their annotation system is the adoption of a common genomic nomenclature. Homologous genes in different organisms have the same gene symbol, facilitating their automated search and association. Hence, the initial gene list was used as the entry to construct a proteostasis profile for thousands of prokaryotic species. These profiles were extracted from the UniProt Knowledgebase (The UniProt Consortium, 2018) as it contains numerous reference proteomes of fully sequenced species. To focus on taxonomically proximal species with 'good quality' genomic annotations, only species with complete proteome detector (CPD index)

equal to “Standard” were selected. As thousands of Bacteria met that criterion, a random selection was then performed to reduce their number to 250. Filtering process had the constraint to select at least one member for each taxonomic Class. Ribosomal sequences (18S and 16S rRNAs) were retrieved from the ENA repository (Leinonen et al., 2010), which contains sequenced genomes and relative information for a wide range of organisms. Heat shock proteins data were collected from Ensembl and UniProt Knowledgebase (**Fig 1A, step II**).

Pathway analysis - PN semantic profile, based on the Biological Process domain of Gene Ontology (GO-BP) was constructed for each species **(Fig. 1B, step I**). We used the prioritized pathway analysis implemented in the BioInfoMiner genomic interpretation platform (Koutsandreas et al., 2016; https://bioinfominer.com/). BioInfoMiner utilizes enrichment statistics together with non-parametric permutation resampling to correct and prioritize the final set of enriched pathways. Furthermore, it adopts the true path rule (G. Valentini, 2011) on ontological graphs to mitigate annotation pitfalls. For each organism, this semantic processing of gene lists identified a network of biological processes, delineating the semantic tree of proteostasis. We used two criteria to prioritize the biological processes. Hypergeometric p-value was set to 0.05 and terms with higher probabilities were filtered out. Then, the adjusted p-value was determined at 0.05 and if an organism had fewer than one hundred terms satisfying that threshold, the selection was extended to the first hundred terms to keep a similar order of magnitude between list sizes.

Unbiased GO-BP and PN profiles - The GO-BP graphs represent an important omnibus of biological knowledge suitable to provide a genome-wide scaffold for systemic interpretation of experimental results. Yet, inherent inconsistencies regarding the structure and the depth of its branches generate bias that hampers comparative analysis (Gaudet & Dessimoz, 2017). Some graph branches are more expanded than others due to extensive annotation, such as the regulation of biological processes. This results in distorted, descriptive capacity, regarding the degree of semantic specification that each term bears, meaning that the graph branches do not carry the same semantic weight. Furthermore, the depth of genomic annotation differs among species. Different research communities have developed meticulous genomic annotations for model species, emphasizing on specific components of cellular physiology, according to specific, operational or temporal characteristics of each organism (e.g. the developmental processes of Zebrafish (Gaudet & Dessimoz, 2017)). On the other hand, the vast majority of organisms has been sketchily annotated, only through electronic inference based on homology associations (Gaudet et al., 2011). This causes

inconsistencies regarding the semantic network profile across species, and even taxonomically related species could have divergence in annotation coverage (Lobb et al., 2020). All these predicate upon a standardized version of GO, suitable for comparative, evolutionary analyses. The construction of such a standardized graph relies on two metrics of ontological graphs, Information Content (IC) and Semantic value (SV). IC measures the specificity of a term, considering the amount of its descendant nodes (Pesquita et al., 2009). Conceptually, a term with plenty of descendant nodes has low IC value because its semantic content can be further broken down into more specific concepts. On the other hand, graph leaves have the maximum IC value, as they are the final terms in the path from the root. IC is defined as follows:

$$IC_t = -log_2\left(\frac{D_t + 1}{N}\right)$$

where D_t refers to the number of descendants of term t and N is the cardinality of the complete set of terms.

As it considers only the number of its child-terms, IC does not integrate the topological information of a term. For example, all leaf nodes have the same IC, but are located at different depths within the graph, depending on the quality of annotation. SV is a metric proposed to overcome that limitation. It depicts the semantic distance of a term from the root node, considering the information contained in its ancestral plexus (X. Song et al., 2014):

$$SV_t = \sum_{a \in A} \frac{1}{1 + exp(-IC_a^{-1})}$$

A is the set of ancestors of term t. The SV of a term is linearly correlated to the amount of its ancestors. High values point out increased distances from the root or the existence of neatly described semantic regions, where terms have plenty of ancestors compared to other more succinct branches. This causes annotation bias favoring the overpopulated tree branches.

For the purposes of the comparative analysis, we delineated a template of PN semantic graph that would render possible the comparison among species, by trading off between the detail of annotation of semantic branches and the need to minimize annotation bias. We created a standardized version of GO-BP by filtering out very specific terms (high IC) from expanded ontological areas (high SV). Terms exceeding the twentieth (20th) percentiles of IC and SV distributions were trimmed and substituted with their most proximal ancestors, conforming to these thresholds (**Figure 1C**). The initial set of 20,913 annotated GO-BP terms was decreased to 2,000 terms. Then, the PN profile of each organism was projected on the standardized version of GO-BP, modifying its content and structure with the terms selected for the comparative analysis (**Figure 1B, step II**).

Comparative analysis of the PN profiles - To quantify the major differences between the proteostasis circuitries of the investigated species, a comparative analysis of their PN profiles was performed. This comparison exploited the semantic framework provided by the hierarchical structure of GO-BP, in order to calculate the group-wise semantic similarities of GO-BP terms lists. Semantic comparison estimates the closeness of two ontological terms, and it is based on the topological relevance of their ancestors (pairwise measures). However, this comparison can be generalized to sets of terms using averaging measures. Hence, a unified pairwise similarity matrix was constructed for all GO-BP terms and the average scores were calculated for each pair of species. To avoid bias of specific pairwise measures, we averaged the similarity of two terms through three validated approaches: Resnik (Resnik, 1999), XGraSm (Mazandu et al., 2016) and AggregateIC (X. Song et al., 2014). The semantic similarity of two species was calculated based on the average best matches formula (Mazandu et al., 2016):

$$SemSim(O_1, O_2) = \frac{\sum_{t_n \in GO_1}^{N} max[PairSemSim(t_n, GO_2)] + \sum_{t_m \in GO_2}^{M} max[PairSemSim(t_m, GO_1)]}{|N| + |M|}$$

where $GO_1$and GO_2 are the GO-BP sets of terms for the pair of species compared and *N*, *M* their cardinalities. Each sum function in the numerator refers to one of the two GO-BP sets of terms and aggregates the maximum similarities of its terms with the other set of terms. The aggregation of best matches between these two lists is averaged by dividing it with the sum of their cardinalities. The final distance matrix was defined by subtracting similarity scores by one (**Figure 1B, step II**), and the phylogenetic dendrogram was generated, using an agglomerative clustering process based on the Ward's minimum variance method (Ward, 1963).

Phylogenetic analysis – Gene sequences of 16S and 18S rRNAs were used to construct the reference phylogenetic tree, as they traditionally portray the evolutionary proximities of species. The ClustalW algorithm (Sievers & Higgins, 2018) was used to quantify the pairwise distances and construct the final distance matrix, based on an *ad hoc* multiple sequence alignment (MSA). Furthermore, amino acid sequences of heat shock protein families (HSP40 and HSP70) were analyzed to examine their potential as surrogate evolutionary markers. Consensus sequences for the HSP proteins were established. Heat shock proteins of the same molecular weight could vary significantly even in the same organism. Each protein class consists of different members, which encompass some identical functional domains, but other additional components or their tertiary structure might be different. For instance, the human genome encodes 13 proteins of the HSP70 family and around 50 members of HSP40, which are divided in three main subfamilies (Kampinga et al., 2009). Members of the same HSP family were clustered to a consensus sequence pattern for each organism. Starting from the

whole set of amino acid sequences, fragments were filtered out. The trimmed part fed the CD-HIT clustering algorithm with similarity threshold to 95% (Li & Godzik, 2006). CD-HIT keeps the longest sequence as the representative of each cluster, conserving as more information as possible for each one. If the output included more than one clusters, then an extra step was performed, by constructing their multiple sequence alignment (MSA) with ClustalW and the respective hidden Markov model (HMM) with the HMMER3 hmmbuild algorithm (Finn et al., 2011). The final consensus sequence was generated using the hmmemit function of HMMER3. ClustalW was used to calculate the distance matrices of consensus sequences, similar to the case of ribosomal sequences. In order to compare the phylogenetic dendrograms and evaluate their discriminative power, as well as their efficiency to reproduce well-shaped taxonomic clusters, organisms were projected on a two-dimensional plot, based on their distances. Specifically, the distance matrix of each phylogenetic approach was transformed into a two-dimensional orthogonal space, using the Multidimensional Scaling (MDS) technique (Borg & Groenen, 2005). MDS executes non-linear dimensionality reduction, projecting the data on a new orthogonal space, where distances among samples converge to the initial values, under a relative tolerance of cost function. The generated scatter plots illustrate the adjacency of species groups, indicating the divergence of each criterion through evolution.

Investigation of proteostasis core components – Enriched GO-BP were clustered hierarchically using the standardized version of the ontology, in order to reveal the core components of PN among the examined species. Resnik's measure (Resnik, 1999) was used to substitute each clustered pair of terms with their lowest common ancestor. Similarity threshold was set to 0.175, as lower values produce very generic clusters with trivial or overly broad semantic description, a technically non-informative result. The final set of 56 generic biological processes contained all the enriched terms used for the comparative analysis. We then performed a feature selection procedure based on the Gain Ratio (Han et al., 2011) to identify the components of proteostasis supporting the strongest taxonomic discrimination between the three domains. First, we explored the classifying power of biological processes associated with eubacteria but not Archaea, targeting to separate the Prokaryotes. Then, the same process was applied for the separation of Eukaryotes and Prokaryotes, looking for features associated with the majority of Eukaryotes. The resulting pseudo-decision tree integrates high Gain Ratio mechanisms that could comprise informative features for the separation of kingdoms. A certain part of PN components was annotated for the vast majority of organisms (>90%). That group was called "common components", while all the rest comprised the "different components" set. Phylogenetic analysis was performed in order to scrutinize their contribution as

regards the taxonomic separability of the complete PN profile. Particularly, we repeated the aforementioned comparative analysis for the different sets of components, with the use of semantic similarity measures, as well as the projection of organisms on the two-dimensional space with MDS.

Evaluation of the classification performances of other evolutionary conserved mechanisms - GO-BP annotation was exploited to inspect the clustering performance of another 20 biological processes, conserved across all analyzed species (Figure S10-S29). Gene lists were retrieved from GO-BP for each organism and biological process. In order to mitigate bias relevant to the cardinality of the different gene lists, these lists were limited to a size (threshold) proportional to that of the corresponding proteostasis gene list. For initially larger gene lists than that of proteostasis, 100 random gene sets with size equal to the respective threshold were generated, in order to neutralize bias attributable to sampling. Pathway analysis of these sets of genes, mapped to the examined biological processes, was implemented to obtain a semantic network profile (or an average profile from the 100 random samplings) for each organism. Comparative analysis was used to calculate the final semantic distance matrix and its phylogenetic dendrogram (averaging organisms' pairwise distances from the 100 distance matrices, if applicable). Each phylogenetic tree was divided in three clusters, aiming to evaluate whether the examined biological process could reproduce the three taxonomic domains. Their efficiency was quantified with Homogeneity and Silhouette scores.

PN contribution to other evolutionary conserved mechanisms - To derive semantic network profiles limited to the non-PN component of these biological processes and therefore measure the performance of PN modules for taxonomic classification, PN components were excluded from the gene lists while keeping the lists sizes proportional. The comparative analysis was also performed on these shortened lists.

Gain Ratio, Homogeneity Score and Silhouette Score – Gain Ratio (Han et al., 2011) is used as an feature selection measure in data mining. Given a dataset *A*, with samples belonging to a set of classes $C = \{c_1, c_2, \dots, c_N\}$ and *D* a subset of *A,* the entropy of the classes' distribution in subset *D* is defined as follows:

$$H(D) = -\sum_{i=1}^{|C_D|} \frac{|C_{i,D}|}{|D|} log\left(\frac{|C_{i,D}|}{|D|}\right)$$

where $|C_{i,D}|$ is the amount of samples in partition *D*, which belong to class C_i. $H(D)$ quantifies the expected information, to classify correctly a sample a_i in *D*. This entropy is maximized when each

sample belongs to a different class. During the training process of decision trees, discrete or continuous variables (features) are examined to separate samples, with respect to their classes. Assume that a feature *F* is selected to separate further the samples in *D*, producing *M* groups. Then, the entropy of *D* is re-calculated, taking into account that partitioning:

$$H_F(D) = \sum_{j=1}^{|M|} \frac{|M_j|}{|D|} \times H(M_j)$$

which is the weighted mean entropy of the derived *M* subsets. Information Gain of feature *F* is equal to the reduction of entropy after the split:

$$InfoGain(F) = H(D) - H_F(D)$$

Another useful measure is the Split Entropy, which is defined as the derived uncertainty due to the partitioning of samples:

$$Hsplit_F(D) = -\sum_{j=1}^{|M|} \frac{|M_j|}{|D|} \times log\left(\frac{|M_j|}{|D|}\right)$$

Split Entropy value increases in function with the amount of the produced subsets and it is maximized when feature *F* creates a novel branch for each sample. It could function as a penalty factor, in order to avoid the selection of features, which tend to segregate the dataset into numerous clusters. Gain Ratio uses that factor to normalize the Information Gain, providing an unbiased measure of splitting information:

$$GainRatio(F) = \frac{InfoGain(F)}{Hsplit_F(D)}$$

Homogeneity Score (HS) (Rosenberg & Hirschberg, 2007) and Silhouette Score (SS) (Rousseeuw, 1987) evaluate specific properties of an unsupervised clustering outcome, given the true classes of the data samples and their pairwise distances. Assume that the samples of dataset *A* are being classified into a set of clusters $K = \{k_1, k_2, \dots, k_m\}$, using an unsupervised clustering algorithm. The amount of samples of class i which are assigned to cluster j is denoted as $|A_{ij}|$. HS evaluates the quality of the derived *K* clusters to contain objects belonging to a unique classes, by measuring the conditional entropy of the classes' distribution given the proposed clustering:

$$H(A|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{|A_{ck}|}{|A|} log\left(\frac{|A_{ck}|}{|A_c|}\right)$$

Formally, $H(A|K)$ quantifies the uncertainty about the distribution of samples in the set of C classes, given the *K* clusters. If each cluster k is homogenous (i.e. containing samples from only one class), then the conditional entropy $H(A|K)$ is equal to zero. The conditional entropy is maximum (equal to $H(A)$ - the entropy of the classes' distribution in *A*) when the proposed *K*-clustering does not

provide any information about the real classification. Homogeneity score is defined as:

$$HS = 1 - \frac{H(A|K)}{H(A)}$$

Using both the normalized entropy and the subtraction, HS is bounded in the range [0, 1] and its desirable value is equal to 1. Silhouette score is a measure of cluster cohesion membership, as it quantifies the trade-off between intra- and inter-distances of the derived clusters. Given the aforementioned *K*-clustering, for each sample *i* assigned to cluster k_n, the factors $f_1(i)$ and $f_2(i)$ are defined as follow:

$$f_1(i) = \frac{1}{|k_n| - 1} \sum_{j \in k_n, j \neq i}^{|k_n - 1|} d(i,j)$$

$$f_2(i) = \min_{m \neq n} \left[\frac{1}{|k_m|} \sum_{j \in k_m}^{|k_m|} d(i,j) \right]$$

where $d(i,j)$ is the distance between samples *i* and *j*. The factor $f_1(i)$ is the average distance of sample *i* to the other samples in cluster k_n and indicates the merit of the assignment to that cluster. The factor $f_2(i)$ is the minimum average distance of sample *i* to all samples in any other cluster, apart from cluster k_n. Namely $f_2(i)$ measures the inter-distance of sample *i* to its neighboring cluster. The Silhouette coefficient of sample *i* is defined as:

$$SS_i = \begin{cases} \dfrac{f_2(i) - f_1(i)}{\max\{f_1(i), f_2(i)\}} & if\ |k_n| > 1 \\ 0 & if\ |k_n| = 1 \end{cases}$$

SS_i ranges from -1 to +1, where a high value implies that *i* is well-located in its group and far from the other clusters (without considering if it is correctly classified). The mean Silhouette over all data samples measures the average cohesion of the derived clusters:

$$SS = \frac{1}{|A|} \sum_{i \in A}^{|A|} ss(i)$$

Acknowledgements

This work was funded by grants from INCa (PLBio), FRM (DEQ20180339169) and ANR (ERAAT) to EC; from INSERM (International Research Project – TUPRIC), EU H2020 MSCA RISE-734749 (INSPIRED) to AC and EC; from ANR (ANR-15-CE12-0003-01) and FRM (DBF20160635724) to BF and ELIXIR-GR (MIS 5002780) to TK.

Author Contributions

AC and TK developed and implemented the computational approach, BF and EC provided the biological conceptual framework to the analyses. All authors worked on the manuscript.

Author Contributions

EC is a founder of Cell Stress Discoveries Ltd. AC is founder and CEO of e-NIOS Applications PC.

Code and Data Availability

The main part of code written in support of this manuscript is publicly available on a GitHub repository at https://github.com/thodk/proteostasis_imprinting_across_evolution . Some tasks were performed using the software of e-NIOS Applications PC, so the respective code is not available. All the results and the data used to produce them, are available in a figshare collection at 10.6084/m9.figshare.c.5427279 .

References

Balch, W. E., Morimoto, R. I., Dillin, A., & Kelly, J. W. (2008). Adapting Proteostasis for Disease Intervention. *Science*, *319*(5865), 916. https://doi.org/10.1126/science.1141448

Borg, I., & Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. https://doi.org/10.1007/978-1-4757-2711-1

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2018). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29–W37. PubMed. https://doi.org/10.1093/nar/gkr367

G. Valentini. (2011). True Path Rule Hierarchical Ensembles for Genome-Wide Gene Function Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(3), 832–847. https://doi.org/10.1109/TCBB.2010.38

Gaudet, P., & Dessimoz, C. (2017). Gene Ontology: Pitfalls, Biases, and Remedies. In C. Dessimoz & N. Škunca (Eds.), *The Gene Ontology Handbook* (pp. 189–205). Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_14

Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, *12*(5), 449–462. PubMed. https://doi.org/10.1093/bib/bbr042

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. PubMed. https://doi.org/10.1093/nar/gky1055

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, *319*(6055), 618–618. https://doi.org/10.1038/319618a0

Gur, E., Katz, C., & Ron, E. Z. (2005). All three J-domain proteins of the Escherichia coli DnaK chaperone machinery are DNA binding proteins. *FEBS Letters*, *579*(9), 1935–1939. https://doi.org/10.1016/j.febslet.2005.01.084

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway

knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. PubMed. https://doi.org/10.1093/nar/gkz1031

Kampinga, H. H., Hageman, J., Vos, M. J., Kubota, H., Tanguay, R. M., Bruford, E. A., Cheetham, M. E., Chen, B., & Hightower, L. E. (2009). Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress & Chaperones*, *14*(1), 105–111. PubMed. https://doi.org/10.1007/s12192-008-0068-7

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, *28*(1), 27–30.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093. https://doi.org/10.1093/bib/bbx085

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation*, *2011*, bar030–bar030. PubMed. https://doi.org/10.1093/database/bar030

Koutsandreas, T., Binenbaum, I., Pilalis, E., Valavanis, I., Papadodima, O., & Chatziioannou, A. (2016). Analyzing and Visualizing Genomic Complexity for the Derivation of the Emergent Molecular Networks. *Int. J. Monit. Surveill. Technol. Res.*, *4*(2), 30–49. https://doi.org/10.4018/IJMSTR.2016040103

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., … Cochrane, G. (2010). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(suppl_1), D28–D31. https://doi.org/10.1093/nar/gkq967

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G., & Doxey, A. C. (2020). An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, *6*(3). https://doi.org/10.1099/mgen.0.000341

Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M., & Mulder, N. J. (2016). A-DaGO-Fun: An adaptable Gene Ontology semantic similarity-based functional analysis tool. *Bioinformatics (Oxford, England)*, *32*(3), 477–479. PubMed. https://doi.org/10.1093/bioinformatics/btv590

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, *5*(7), e1000443. https://doi.org/10.1371/journal.pcbi.1000443

Phillips, A. M., Gonzalez, L. O., Nekongo, E. E., Ponomarenko, A. I., McHugh, S. M., Butty, V. L., Levine, S. S., Lin, Y.-S., Mirny, L. A., & Shoulders, M. D. (2017). Host proteostasis modulates influenza evolution. *ELife*, *6*, e28652. https://doi.org/10.7554/eLife.28652

Powers, E. T., & Balch, W. E. (2013). Diversity in the origins of proteostasis networks—A driver for protein function in evolution. *Nature Reviews. Molecular Cell Biology*, *14*(4), 237–248. PubMed. https://doi.org/10.1038/nrm3542

Powers, E. T., Powers, D. L., & Gierasch, L. M. (2012). FoldEco: A model for proteostasis in E. coli. *Cell Reports*, *1*(3), 265–276. PubMed. https://doi.org/10.1016/j.celrep.2012.02.011

Rebeaud ME, Mallik S, Goloubinoff P, Tawfik DS. On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. Proc Natl Acad Sci U S A. 2021 May 25;118(21):e2020885118.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 410–420. https://www.aclweb.org/anthology/D07-1043

Roth, D. M., & Balch, W. E. (2011). Modeling general proteostasis: Proteome balance in health and disease. *Current Opinion in Cell Biology*, *23*(2), 126–134. PubMed. https://doi.org/10.1016/j.ceb.2010.11.001

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shemesh, N., Jubran, J., Dror, S., Simonovsky, E., Basha, O., Argov, C., Hekselman, I., Abu-Qarn, M., Vinogradov, E., Mauer, O., Tiago, T., Carra, S., Ben-Zvi, A., & Yeger-Lotem, E. (2021). The landscape of molecular chaperones across human tissues reveals a layered architecture of

core and variable chaperones. *Nature Communications*, *12*(1), 2180. https://doi.org/10.1038/s41467-021-22369-9

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290

Smit, S., Widmann, J., & Knight, R. (2007). Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Research*, *35*(10), 3339–3354. https://doi.org/10.1093/nar/gkm101

The UniProt Consortium. (2018). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Vonk, W. I. M., Rainbolt, T. K., Dolan, P. T., Webb, A. E., Brunet, A., & Frydman, J. (2020). Differentiation Drives Widespread Rewiring of the Neural Stem Cell Chaperone Network. *Molecular Cell*, *78*(2), 329-345.e9. https://doi.org/10.1016/j.molcel.2020.03.009

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wiseman, R. L., Powers, E. T., Buxbaum, J. N., Kelly, J. W., & Balch, W. E. (2007). An Adaptable Standard for Protein Export from the Endoplasmic Reticulum. *Cell*, *131*(4), 809–821. https://doi.org/10.1016/j.cell.2007.10.025

X. Song, L. Li, P. K. Srimani, P. S. Yu, & J. Z. Wang. (2014). Measure the Semantic Similarity of GO Terms Using Aggregate Information Content. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(3), 468–476. https://doi.org/10.1109/TCBB.2013.176

Legends to figures

Figure 1: **Schematic analytical workflow** – **(A)** Process of data acquisition. Proteostasis-related gene lists were constructed manually for model organisms (step I). Then, public databases were used to collect genomic homologies to expand organisms' collection and concentrate specific data for rRNA, HSP40 and HSP70 sequences. Only organisms with ample annotation were selected in the final set (step II). **(B)** Analysis workflow. Gene lists were translated into networks of biological processes through pathway analysis of GO-BP corpus (step I). An unbiased version of GO was constructed to remove potential annotation bias from the results. Finally, semantic analysis was performed to calculate the differences among proteostasis-related networks (step II). **(C)** Construction of the unbiased GO-BP graph: Terms residing low in the branches of the graph bear high IC, meaning they are really specific with the leaves of ontological graph bearing maximum IC (red and blue nodes).At the same time, some of them are located in deep, densely populated branches (red nodes), reflecting the fact that these processes are more extensively studied than others, producing great semantic values. This imbalance in knowledge representation is a source of annotation bias, which was neutralized by filtering out terms in voluminous ontological regions (elongated and abundant branches), with high information content (red nodes).

Figure 2: **Establishment of PN-based phylogenetic trees** – **(A)** Two-dimensional representation of organisms in function with their taxonomy. The derived evolutionary distance matrix of each criterion (rRNA, proteostasis and HSPs), was transformed into a 2-dimensional orthogonal space through the Multidimensional Scaling (MDS) algorithm, reflecting the two larger, dimensions of observed variation (termed eigen-components). Similarity distances of each organism from the centroid of the single class problem are projected in those exploratory scatter plots. **(B)** Pearson correlation of pairwise distances of rRNA sequences with the other three measures. Correlation is unbiased from taxonomic domain sizes, as we used 80 randomly selected species from each domain for the calculation.

Figure 3: PN composition evolution throughout various phyla – **(A)** The association matrix of proteostasis components with organisms set. These components were revealed performing a semantic clustering of enriched terms on the GO-BO graph, using the measure of Rensik. The red section refers to the mechanisms associated with more than 90% of species and it is defined as "common components". **(B)** A pseudo decision tree indicates the most informative PN components to separate Eukaryotes from Prokaryotes and Bacteria from Archaea. The first batch refers to

mechanisms associated mainly with Eukaryotes and not the prokaryotic world, while the latter one includes mechanisms related with Bacteria and not Archaea. The quantification of their importance was achieved with the Gain Ratio score. **(C)** Two-dimensional representation of organisms based on proteostasis profiles. Common and different mechanistic components were used separately to investigate their contribution to taxonomic domain separation.

Figure 4: **PN contribution to other evolutionary conserved mechanisms** – Homogeneity score (**A, red**) refers to the separating ability of each biological process regarding the three main taxonomic domains, through the respective semantic network, derived from the pathway analysis of related genes. Silhouette score (**B, green**) indicates the degree of cohesion of cluster inference, by measuring the trade-off between intra- and inter-distances of each cluster member. Bars display these scores for the entire machinery of each process whereas the solid lines illustrates the same scores calculated after the removal of proteostasis related components from the gene machinery of each process.

Figure 1

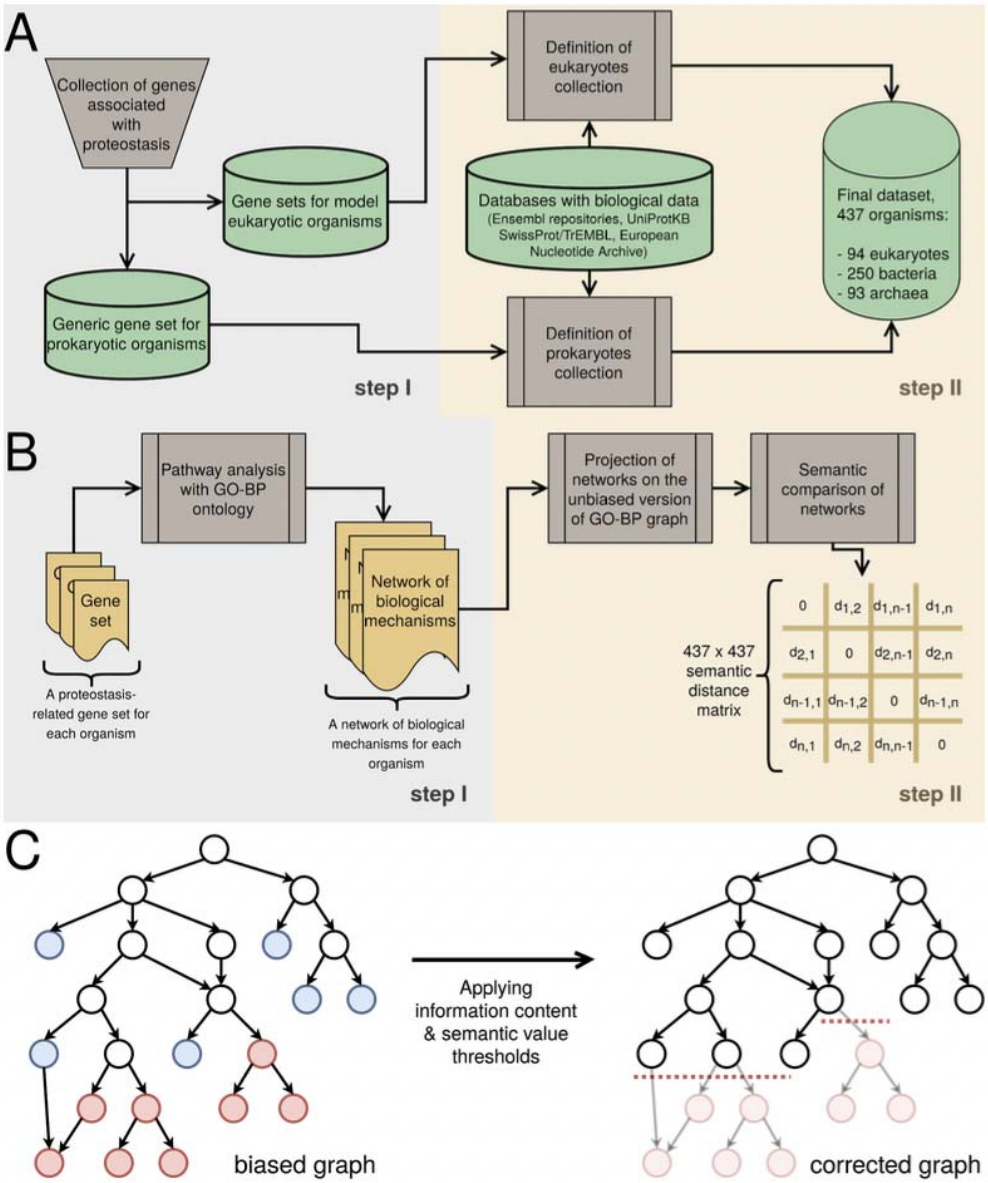

A
Collection of genes associated with proteostasis
Gene sets for model eukaryotic organisms
Generic gene set for prokaryotic organisms
step I
Definition of eukaryotes collection
Databases with biological data (Ensembl repositories, UniProtKB SwissProt/TrEMBL, European Nucleotide Archive)
Definition of prokaryotes collection
Final dataset, 437 organisms:
- 94 eukaryotes
- 250 bacteria
- 93 archaea
step II
B
Pathway analysis with GO-BP ontology
Gene set
A proteostasis-related gene set for each organism
Network of biological mechanisms
A network of biological mechanisms for each organism
step I
Projection of networks on the unbiased version of GO-BP graph
Semantic comparison of networks
437 x 437 semantic distance matrix
0 d1,2 d1,n-1 d1,n
d2,1 0 d2,n-1 d2,n
dn-1,1 dn-1,2 0 dn-1,n
dn,1 dn,2 dn,n-1 0
step II
C
Applying information content & semantic value thresholds
biased graph
corrected graph

Figure 2

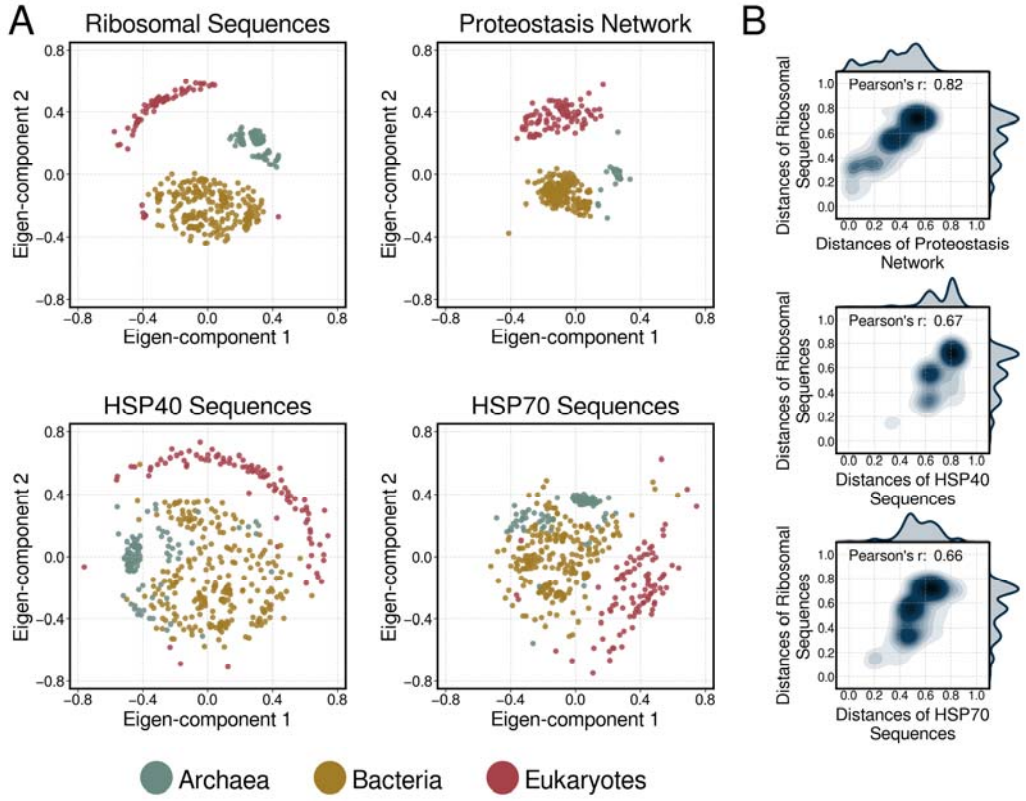

A
Ribosomal Sequences
Proteostasis Network
HSP40 Sequences
HSP70 Sequences
Eigen-component 1
Eigen-component 2
B
Pearson's r: 0.82
Distances of Ribosomal Sequences
Distances of Proteostasis Network
Pearson's r: 0.67
Distances of HSP40 Sequences
Pearson's r: 0.66
Distances of HSP70 Sequences
Archaea
Bacteria
Eukaryotes

Figure 3

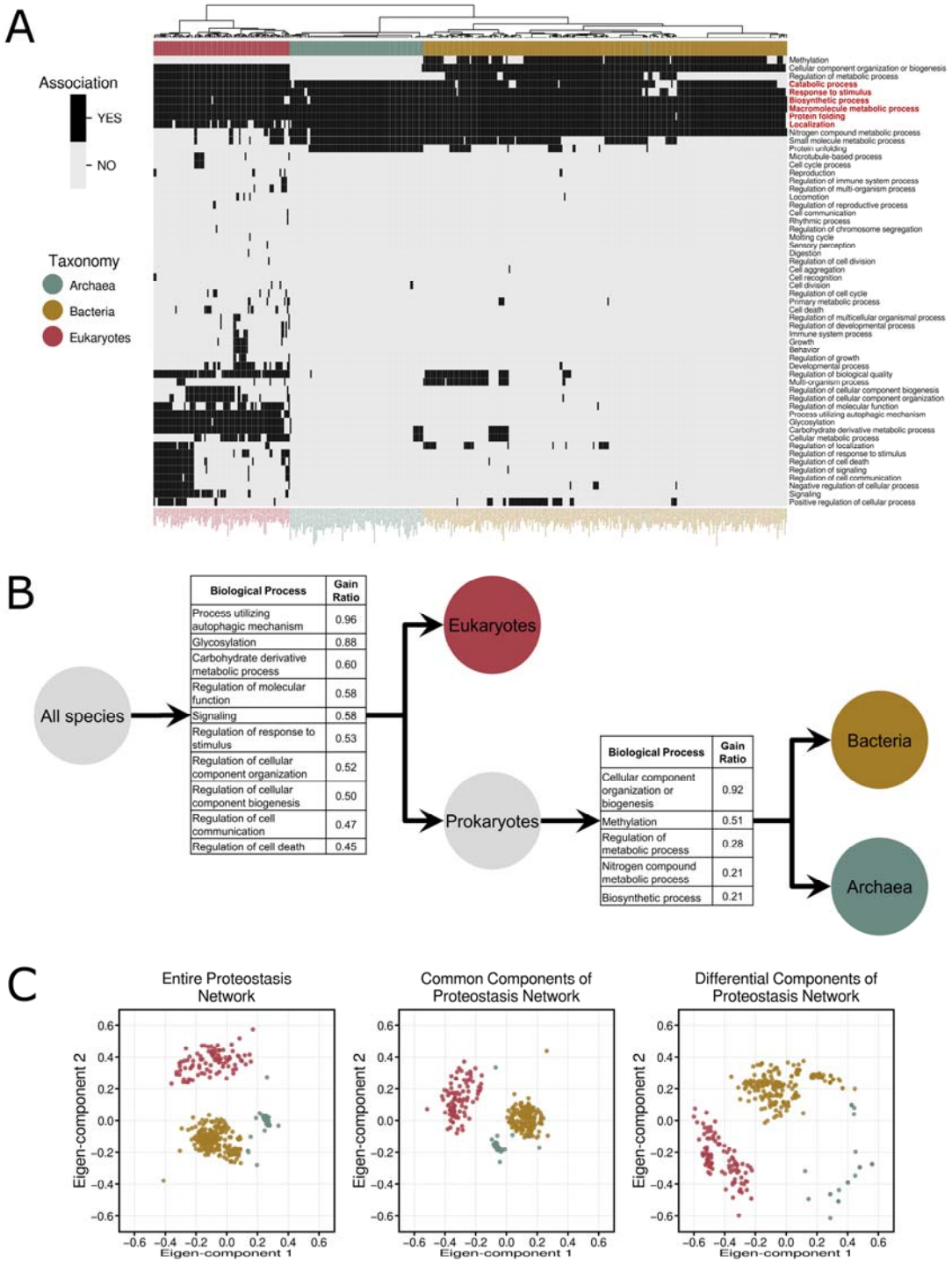
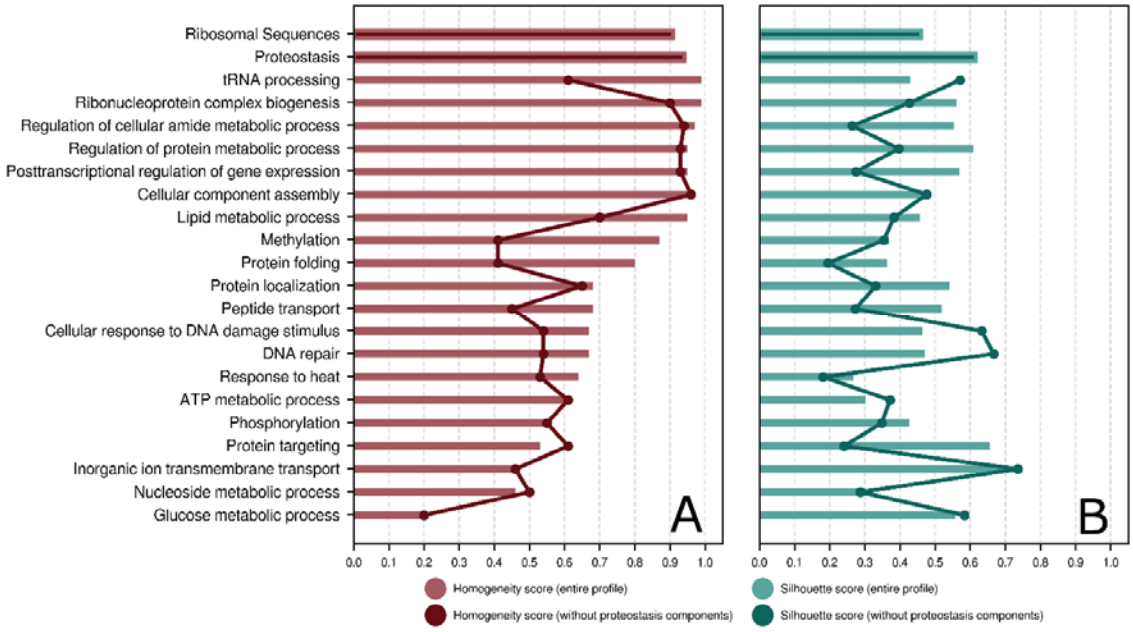

A
Association
YES
NO
Taxonomy
Archaea
Bacteria
Eukaryotes
Methylation
Cellular component organization or biogenesis
Regulation of metabolic process
Catabolic process
Response to stimulus
Biosynthetic process
Macromolecule metabolic process
Protein folding
Localization
Nitrogen compound metabolic process
Small molecule metabolic process
Protein unfolding
Microtubule-based process
Cell cycle process
Reproduction
Regulation of immune system process
Regulation of multi-organism process
Locomotion
Regulation of reproductive process
Cell communication
Rhythmic process
Regulation of chromosome segregation
Molting cycle
Sensory perception
Digestion
Regulation of cell division
Cell aggregation
Cell recognition
Cell division
Regulation of cell cycle
Primary metabolic process
Cell death
Regulation of multicellular organismal process
Regulation of developmental process
Immune system process
Growth
Behavior
Regulation of growth
Developmental process
Regulation of biological quality
Multi-organism process
Regulation of cellular component biogenesis
Regulation of cellular component organization
Regulation of molecular function
Process utilizing autophagic mechanism
Glycosylation
Carbohydrate derivative metabolic process
Cellular metabolic process
Regulation of localization
Regulation of response to stimulus
Regulation of cell death
Regulation of signaling
Regulation of cell communication
Negative regulation of cellular process
Signaling
Positive regulation of cellular process
B
All species
Biological Process | Gain Ratio
Process utilizing autophagic mechanism | 0.96
Glycosylation | 0.88
Carbohydrate derivative metabolic process | 0.60
Regulation of molecular function | 0.58
Signaling | 0.58
Regulation of response to stimulus | 0.53
Regulation of cellular component organization | 0.52
Regulation of cellular component biogenesis | 0.50
Regulation of cell communication | 0.47
Regulation of cell death | 0.45
Eukaryotes
Prokaryotes
Biological Process | Gain Ratio
Cellular component organization or biogenesis | 0.92
Methylation | 0.51
Regulation of metabolic process | 0.28
Nitrogen compound metabolic process | 0.21
Biosynthetic process | 0.21
Bacteria
Archaea
C
Entire Proteostasis Network
Common Components of Proteostasis Network
Differential Components of Proteostasis Network
Eigen-component 2
Eigen-component 1
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 4


Ribosomal Sequences
Proteostasis
tRNA processing
Ribonucleoprotein complex biogenesis
Regulation of cellular amide metabolic process
Regulation of protein metabolic process
Posttranscriptional regulation of gene expression
Cellular component assembly
Lipid metabolic process
Methylation
Protein folding
Protein localization
Peptide transport
Cellular response to DNA damage stimulus
DNA repair
Response to heat
ATP metabolic process
Phosphorylation
Protein targeting
Inorganic ion transmembrane transport
Nucleoside metabolic process
Glucose metabolic process
A
B
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Homogeneity score (entire profile)
Homogeneity score (without proteostasis components)
Silhouette score (entire profile)
Silhouette score (without proteostasis components)