Protein homeostasis imprinting across evolution

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

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

Classification: Biophysics and Computational biology

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

****Corresponding Authors:*** EC (eric.chevet@inserm.fr), AC (achatzi@eie.gr) and BF (brice.felden@univ-rennes1.fr).

Abstract

The control of 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 has evolved to adjust the protein biogenesis needs with environmental constraints. Herein, we describe an approach to evaluate proteostasis during evolution, and show that the proteostasis network (PN) represents a reliable metric to i) deconvolute the life forms into Archaea, Bacteria and Eukaryotes and ii) assess the evolution rates among species, without the need for rRNA sequences. This method for phylogenetic comparison relies on the use of semantic graphs to evaluate PN complexity. This stands as a novel strategy for taxonomic classification, based on the analysis of 94 Eukaryotes, 250 Bacteria and 93 Archaea genome sequences. A functional analysis was used as a powerful phylogenetic tool that echoes with species complexity. It provides information about species divergence and indicates the taxonomic clades that evolved faster than others did. Phyla-specific PN were identified, suggesting that PN complexity correlates with the grade of evolution the species have reached. Individual components, however, such as the heat shock proteins (HSP) do not accurately mark evolution. We analyzed gene conservation, gain or loss that occurred throughout PN evolution, which reveals a dichotomy within the PN conserved parts (e.g. chaperones), but also with species-specific modules. Since the PN is implicated in cell fitness, aging control and the onset of several diseases, it could be used as a metric to tackle gain-of-functions mechanisms and their biological impact.

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. From lower Bacteria to humans, the molecular components that control proteostasis (i.e. the proteostasis network - PN) include protein synthesis, co/post-translational protein folding, quality control, degradation, as well as adaptive signaling to proteostasis imbalance (Balch et al., 2008). From Prokaryotes to Eukaryotes, the PN was subjected to evolutionary pressures for each organism to cope with intrinsic and extrinsic demands. Evolution is shaped by i) genome complexity, ii) post-translational modifications, iii) subcellular compartmentalization, and iv) the emergence of multicellular organisms and cell differentiation (e.g. cells possessing high protein synthesis yields and secretion requiring a robust PN). Also, for the eukaryotic cells, exhibiting temporal variations of protein expression (e.g. neurons or endocrine cells), a powerful PN is also vital. Each of these constraints increased the needs for updated and adaptive mechanisms to ensure protein homeostasis (Roth & Balch, 2011). As such, the PN was subjected to selection pressures specific to each organism. For instance, compartment-specific proteostasis control machineries in the cytosol, the endoplasmic reticulum (ER), or the mitochondria were required for the eukaryotic cells, whereas dedicated systems arose in the periplasms of Gram-negative Bacteria (Powers & Balch, 2013).

Even though studies have reported computational models of proteostasis in Bacteria (e.g. *E. coli* (Powers et al., 2012)) and in Eukaryotes (Wiseman et al., 2007), an overall layout of proteostasis evolution is lacking, that could lead to applied perspectives. Using an approach evaluating the functions of specific biological modules (i.e. list of genes/proteins or associated biological terms into organized classes of related genes/proteins or biology) during evolution, we assessed how proteostasis evolved among the phylogeny. The proteostasis-based, evolutionary maps performed well *vs*. the gold-standard ribosomal RNA (rRNA) sequence-based evolution metric (Gilbert, 1986; Smit et al., 2007). Specifically, they managed to yield a phylogenetic dendrogram, which proposes taxonomic classifications of organisms, based in the topology of their PN, that generally reproduce the established classifications in phyla. Furthermore, species allocation into distinct clusters, populated by members from various phyla unveils shared adaptive responses to environmental cues. We show here that proteostasis, as a whole, is a reliable tool for species partitioning. 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 the maintenance of protein homeostasis is essential to warrant proper proteome functions, it should represent *de facto* a conserved actor during evolution. Our idea was to study how the PN evolves among organisms, taking into account functional differences and commonalities recorded into the topology of their architectures. This could provide a scaffold enabling delineation of a PN-based phylogenetic tree. First, we built the PN composition, based on the gene lists related to proteostasis according to the Gene Ontology Biological Process (GO-BP) annotation. We also used the available rRNA nucleotide sequences and heat shock protein (HSP) amino acid sequences, to carryout phylogenetic analyses on an identical group of organisms, to compare their results with the PN-based tree. An exhaustive analysis of public repositories of biological data (Ensembl (Cunningham et al., 2018), UniprotKB (The UniProt Consortium, 2018), and ENA (Leinonen et al., 2010)) was performed to collect the raw data. Organisms, which met the following criteria, were included in the analysis: 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*) heat shock proteins. Using both manual and automated procedures (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 **Supplementary Files** gene_lists.zip and gobp_terms.zip). A phylogenetic comparison of these functional profiles was performed through the calculation of the semantic PN similarities. Specifically, we used a standardized version of the GO-BP framework in conjunction with semantic operators, to quantify the similarities of the term lists and create a phylogenetic dendrogram by clustering together species with high 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, neutralizing annotation imbalances, which incur 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). Thereby, 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 into an orthogonal 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 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. 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 essentially conserved across evolution. Nevertheless, the pairwise distances among species for rRNA and PN showed unambiguous correlation (**Figure 2B, Figures S5-6**). Concerning the HSP-based 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**). It implies that heat shock proteins, which are individual components of the PN, lack informative power as marker of species evolution.

We next sought to compare the ability of PN to classify species, compared to that of rRNA and HSPs sequences. Specifically, 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 (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 a number of clusters equal to that of the reference taxonomic groups, verifying that species of different lower-level taxonomies share highly similar profiles. rRNA sequences represented the most effective measure. The PN-based classification indicated that Bacteria and Archaea share PN components. On the other hand, the eukaryotic PN was more complex, encompassing variations that

coherently segregated the species, rescuing phyla segregations. HSP-based 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-based species organization provided a much more informative and robust classification at the Phylum level in Eukaryotes, compared to either HSP40 or HSP70 (**Figure S9**).

Proteostasis as a functional metric to trace evolution

Pathway analysis of gene lists related to proteostasis, 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 regulatory functions were enriched in Eukaryotes, such as proteins associated with programmed cell death or signaling pathways. Enrichment in the response to endoplasmic reticulum (ER) homeostasis imbalance in Eukaryotes was effective in all the tested species, pointing out this compartment as a hotspot for proteostasis, as expected. This was evidenced by functions of the PN associated with glycosylation or carbohydrate metabolism. Similarly, enrichment of membranous organelles-associated mechanisms, such as mitochondrial organization or autophagy, was 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 established a PN conservation tree in Eukaryotes, Bacteria and Archaea to identify the commonalities and discrepancies (**Figure 3B**). We used the association of species with each mechanistic cluster, to derive their two-dimensional representations for the whole PN (**Figure 3C**, left), the common PN (**Figure 3C**, middle) and the phylum-specific PN (**Figure 3C**, right). The PN functional profile produced independent sub-groups for each taxonomic domain. As expected, these sub-groups exhibited an increased density when analyzing the common PN (**Figure 3C**, middle), and showed further expansion when analyzing the differential PN (**Figure 3C**, right). Moreover, analysis of the differential PN unveiled a great diversity within each domain, in particular among Archaea, in which proteostasis is defined by the common PN, but variations in the differential components induced great semantic distances. These data point towards the characteristics of the PN in various species, associated with functions specific to each taxonomic domain. In addition, they unveil conservation and divergence within the PN.

Impacts of PN evolution on other functional networks

To evaluate the robustness of our methodology and analyze the impact of the PN onto the evolution of other cellular pathways, we examined other conserved mechanisms by the same procedure. Their efficiency to separate the main taxonomic domains was evaluated using metrics from Data Mining theory (Homogeneity (HS) and Silhouette (SS) scores, see Materials & Methods). Using the same method as described above, we analyzed 20 conserved cellular pathways as defined in the GO-BP domain. This task was performed to examine the contribution of individual PN components in the form of gene sets, also shared by these mechanisms, to their accuracy. The semantic analysis and phylogenetic tree construction were performed based on their whole pathway annotation. Then, PN components were excluded to repeat that process upon their exclusion. The analysis showed that biological processes, which emerge as gain of function during species evolution, or those which evolved through higher functional complexity, segregate well the three super-kingdoms (**Figure 4; S10A-17A**). 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 for the rRNA sequences). Nevertheless, their phylogenies exhibited lower SS values comparing to PN (**Figure 4B**), which means that the produced clusters are sparser (especially within Prokaryotes) and, consequently, the phylogenetic trees contain broader clades. On the other hand, 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-29A**). 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 various taxonomies. In general, all the processes with accurate performance suffered from low Silhouette scores, and some have lost their phylogenetic congruity. Improvements of 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 about 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 control (the proteostasis network – PN). This stands as a novel strategy for taxonomic classification, which assess divergence of the topological similarity of ontological trees, providing species-specific PN profiles rather than analyzing individual sequences. The crux of this new method is the semantic comparison of the inferred graphs and their topologies. Using an ontological framework, the semantic interpretation of these gene/RNA/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 to illustrate the complexity of the PN modular architecture. In addition, our analysis measures species-specific topological differences, and translate those into evolutionary paths according to the adaptive constraints. Here, we show that monitoring PN characteristics provides reliable information about the evolution of living organisms, whereas monitoring its individual components has limited interpretation.

PN succeeded to separate the three main taxonomic domains almost infallibly, performing as accurately as rRNA sequences do. Furthermore, PN performed much better than isolated PN components, such as the heat shock proteins (**Figure 2**). These clues primarily indicate that the proposed new method discloses 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 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, the latter only meaningful on isolated biological mechanisms. A way to solve that issue could be the massive informational integration from vocabularies with better functional annotations, such as the MetaCyc database (Caspi et al., 2017). In addition, the weaker segregating capacity of the PN in the two Prokaryotic kingdoms suggests 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. Other functional modules, considered as 'gain or loss' of adaptive functions, are also connected to the PN. This, in turn, could help identify the role of PN in maintaining those functional changes (**Figure 4**). This observation is in line with the recently reported observation that core-chaperones were already available in prokaryotes ~3 billion years ago, and expanded linearly from archaea to mammals (Rebeaud et al., 2020). The instrumental role of proteostasis as a robust indicator of cellular and organismal adaptations to evolutionary cues, is highlighted by the taxonomic underperformance that the other mechanisms linked to PN exhibit when the PN components are excluded (**Figure 4**). This is evidence for a tight coupling between proteostasis and most of the other 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 act 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 reliable evolution marker, 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 that could use 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 or 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. The inventories of the species-specific gene lists were performed through a supervised multi-phase analytic workflow. Its objective was to mitigate functional knowledge bias, due to the lack of related annotation for different species by projecting species with a standard PN tree (**Fig 1A, step I**). This supervised, proteostasis-related, gene selection for each species 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, Xenopustropicalis, 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, metazoa and plants, (Kinsella et al., 2011)), as they contain multi-layered data of fully sequenced and adequately, annotated organisms, together with appropriate information schemas that link them through genomic homology with different species. 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). Iterating that process, 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 generic gene list was used as the entry to construct a proteostasis profile for thousands of prokaryotic species. These profiles were

extracted from UniProt Knowledgebase (The UniProt Consortium, 2018) as it contains numerous reference proteomes of fully sequenced species, providing sufficient genomic annotation for a big part of them, through manual or electronic curation. 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 was at random, with the constraint to select at least one member for each taxonomic Class level. 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 – A PN semantic profiles based on the Biological Process corpus of Gene Ontology (GO-BP) was constructed for each species **(Fig. 1B, step I**). We used the prioritized pathway analysis of related gene lists 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 according to the GO-BP annotation identifies a network of biological processes, thus delineating the semantic tree of proteostasis. We used two distinct 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 lists 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 (Gaudet & Dessimoz, 2017), generate bias that hampers comparative analysis. 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 describing biological processes 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 important metrics of ontological graphs, Information Content (IC) and Semantic value (SV). IC measures term specificity, considering the amount of its offspring nodes (Pesquita et al., 2009). Conceptually, a term with plenty of descendant nodes has low information content because its semantic content can be further broken down into more specific concepts. On the other hand, graph leaves maximize their information content, 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 size of the complete set of terms in the ontology.

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 information content (the maximum), but are located at different depths within the graph depending on the quality of annotation and the curation effort made in its distinct branches (sub-graphs), IC also ignores the graph complexity of their ancestors.

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 number of its ancestors, high values point out increased distances from the graph 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 evolutionary comparative analysis, we delineated a template of proteostasis semantic graph that would render possible the comparison among species by trading off

between the detail of annotation for the major cellular processes linked to proteostasis and the need to minimize annotation bias in graph corpus and among species. We created a standardized version of GO-BP by filtering out specific terms (high IC) from expanded ontological areas (high SV). Terms exceeding the twentieth (20th) percentiles of IC and SV distributions were trimmed and an iterative procedure substituted them with their most proximal ancestors, conforming to these thresholds (**Figure 1C**). The initial set of 20,913 annotated GO-BP terms was thus decreased to 2,000 terms. Then, the PN profile of each organism was projected on the standardized version of GO, modifying its content and structure with the terms selected for the comparative analysis (**Figure 1B, step II**).

Comparative analysis of the PN profiles– To identify the major differences between the proteostasis circuitries of the investigated organisms, a comparative analysis of their PN profiles was performed. This comparison exploited semantic framework provided by the hierarchical tree structure of the GO-BP ontology. GO-BP is a directed acyclic graph in which each node refers to a semantic term linked with child and/or parent terms. Molecular pathways share semantic similarity if their terms are connected with an ancestor-descendant relation or if they are located near a common ancestral term. Comparison of the PN profiles was performed through a phylogenetic analysis based on the group-wise semantic similarities between GO-BP terms lists. In general, semantic comparison aims to estimate of closeness of two ontological objects (terms), and it is based on the topological relevance of their ancestry in the ontological graph (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 aggregated scores were the calculated for each pair of organisms. 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 organisms 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 organisms compared and *N*, *M* their cardinalities. Each sum function in the numerator refers to one of two GO-BP sets of terms and aggregates the maximum similarities of each 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 the 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 operator with Ward's minimum variance method as objective function (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 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), which are ubiquitous molecules of proteostasis across the different super-kingdoms, 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 HMMER3 hmmbuild algorithm (Finn et al., 2011). The final consensus sequence was calculated using the hmmemit function of HMMER3. ClustalW was used to calculate the distance matrices of consensus sequences, similar to the case of ribosomal sequences. To juxtapose the phylogenetic dendrograms and compare 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 pair-wise 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 pathways were clustered hierarchically according to the standardized version of GO-BP for comparative, evolutionary analysis. 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 pathways used for the comparative analysis. We then performed a feature selection procedure based on the Gain Ratio to identify the components of proteostasis supporting the strongest taxonomic discrimination between the three domains. First, to separate the Prokaryotes, we explored the classifying power of biological processes associated with eubacteria but not Archaea. Then, the same concept was applied for Eukaryotes and Prokaryotes separation 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 super-kingdoms separation. A certain part of P modules of proteostasis has been PN content 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 two-dimensional space with MDS.

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 defines a different class and becomes zero, when all samples belong to a unique 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 correlation 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 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 a 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 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 maximized when the proposed *K*-clustering does not provide any information about the real classification and equal to the entropy of the classes' distribution (same equation as $H(D)$, where *D*=*A*):

$$H(A) = -\sum_{i=1}^{|C|} \frac{|C_i|}{|A|} log\left(\frac{|C_i|}{|A|}\right)$$

Homogeneity score is defined by subtracting the normalized conditional entropy from 1:

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

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:

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

Evaluation of the classification performances of other evolutionary conserved mechanisms - GO-BP corpus was exploited to inspect the clustering performance of the other biological processes recorded across all analyzed species. As their terms are connected with parent-child relations, a representative part was defined and analyzed following the same comparative semantic analysis framework. Gene lists were retrieved from GO-BP corpus 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 length/size proportional to that of the corresponding proteostasis gene list (threshold). For initially larger gene lists than that of proteostasis, 100 random gene sets with size equal to the threshold were generated in order to neutralize bias attributable to sampling. Pathway analysis of these sets of genes mapped to the examined biological process 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 mechanism 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.

Acknowledgements

We thank Drs Harm Kampinga, Marc Aubry and Gérald Peyroche for critical reading and input on our manuscript. This work was funded by grants from INCa (PLBio), FRM (DEQ20180339169) and ANR (ERAAT) to EC; from 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 competing interests

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

References

Balch, W. E., Morimoto, R. I., Dillin, A., & Kelly, J. W. (2008). Adapting Proteostasis for Disease Intervention. *Science*, *319*(5865), 916.

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

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S., Subhraveti, P., & Karp, P. D. (2017). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*, *46*(D1), D633–D639.

Choi, S., Cha, S.-H., & Tappert, C. (2009). A Survey of Binary Similarity and Distance Measures. *J. Syst. Cybern. Inf.*, *8*.

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.

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.

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.

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.

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.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, *319*(6055), 618–618.

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.

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.

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

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.

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.

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.

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.

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).

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.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, *5*(7), e1000443.

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.

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.

Powers, E. T., Powers, D. L., & Gierasch, L. M. (2012). FoldEco: a model for proteostasis in E. coli. *Cell Reports*, *1*(3), 265–276.

Rebeaud M.E., Mallik S., Goloubinoff P., Tawfik D.S. On the evolution of chaperones and co-chaperones and the exponential expansion of proteome complexity. bioRxiv doi: https://doi.org/10.1101/2020.06.08.140319

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.

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.

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.

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145.

Smit, S., Widmann, J., & Knight, R. (2007). Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Research*, *35*(10), 3339–3354.

The UniProt Consortium. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515.

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

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.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244.

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.

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.

Legends to figures

Figure 1: **Schematic analytical workflow** – **(A)** Process of data acquisition. Proteostasis-related gene lists were constructed manually for model organisms. 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. **(B)** Analysis workflow. Gene lists were translated into networks of biological processes through pathway analysis of GO-BP corpus. 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. **(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 main components with organisms set. These components were revealed performing a semantic clustering of enriched pathways on the GO-BO graph. Each pathway was aggregated to more generic terms, ending up to that set of systemic processes. 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

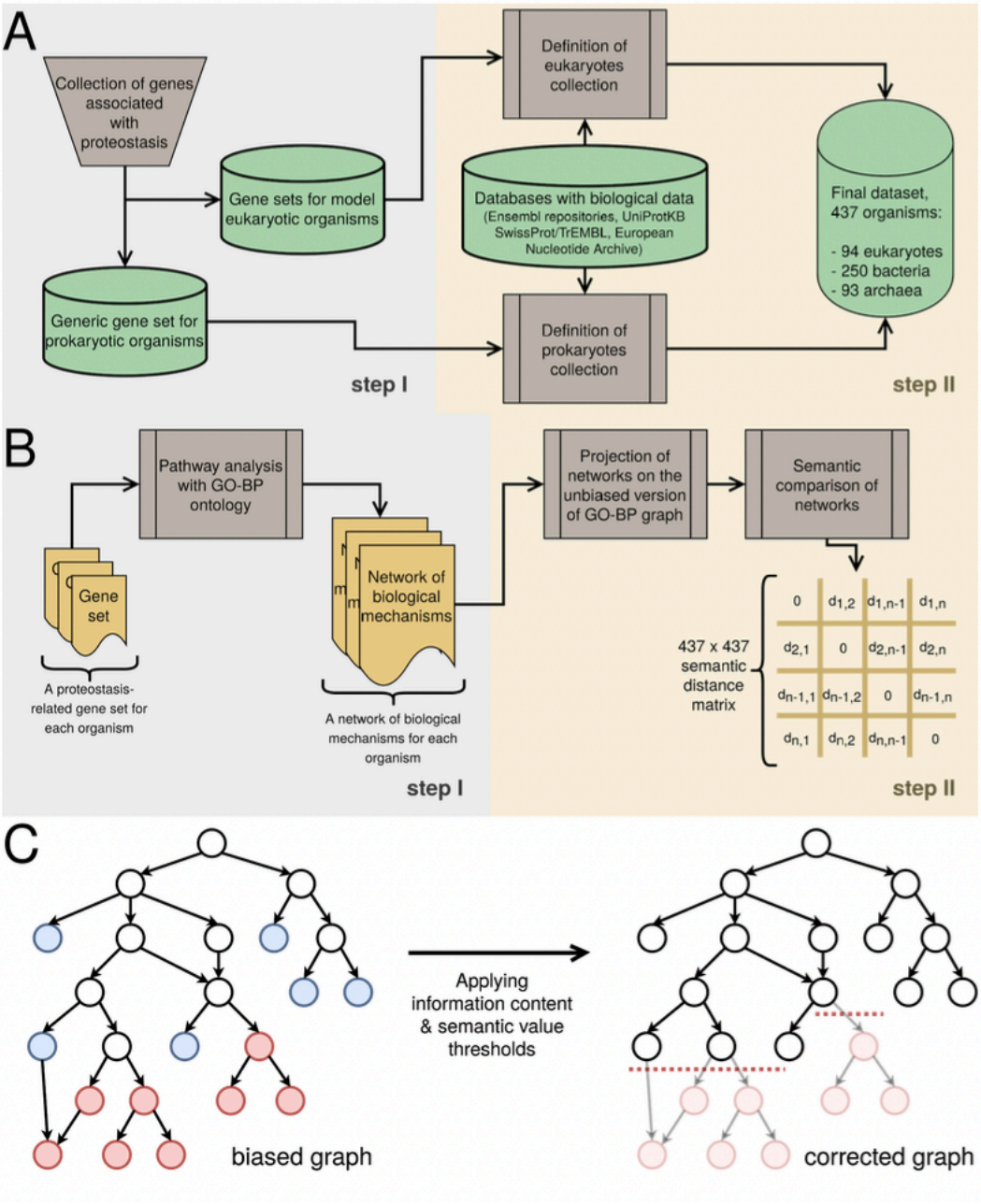

A
Collection of genes associated with proteostasis
Gene sets for model eukaryotic organisms
Generic gene set for prokaryotic organisms
step I
Definition of eukaryotes collection
Databases with biological data (Ensembl repositories, UniProtKB SwissProt/TrEMBL, European Nucleotide Archive)
Definition of 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
biased graph
Applying information content & semantic value thresholds
corrected graph

Figure 2

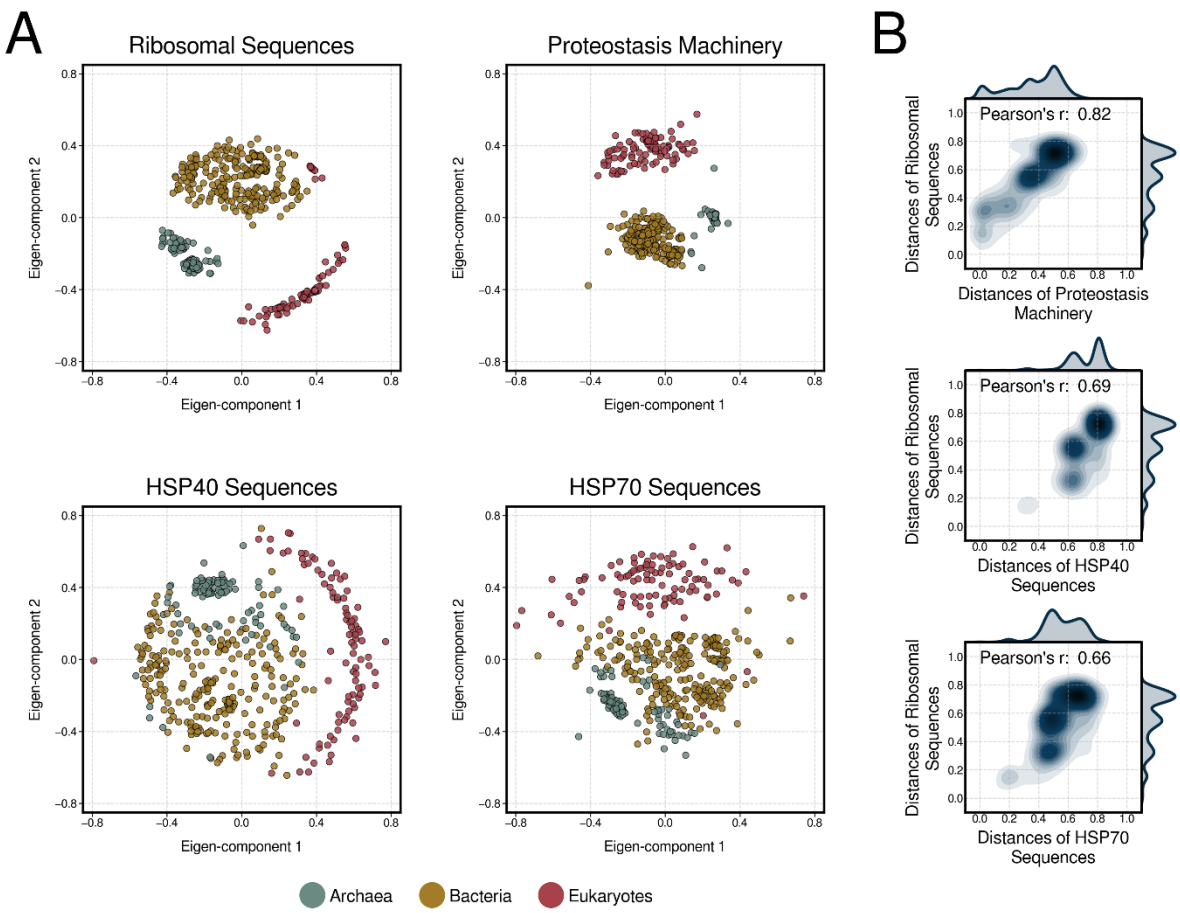
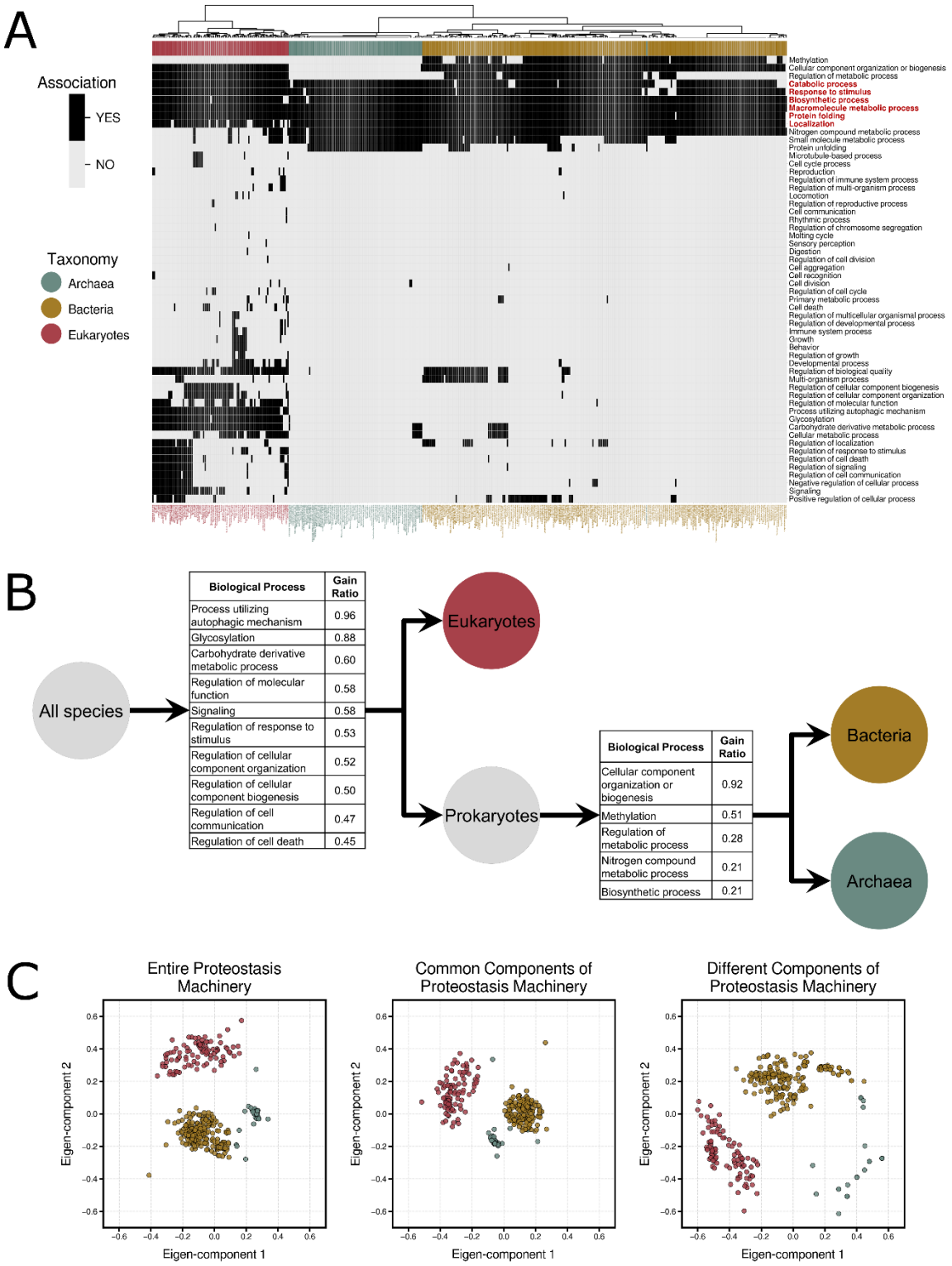

A
Ribosomal Sequences
Proteostasis Machinery
HSP40 Sequences
HSP70 Sequences
Eigen-component 1
Eigen-component 2
B
Pearson's r: 0.82
Distances of Ribosomal Sequences
Distances of Proteostasis Machinery
Pearson's r: 0.69
Distances of HSP40 Sequences
Pearson's r: 0.66
Distances of HSP70 Sequences
Archaea
Bacteria
Eukaryotes

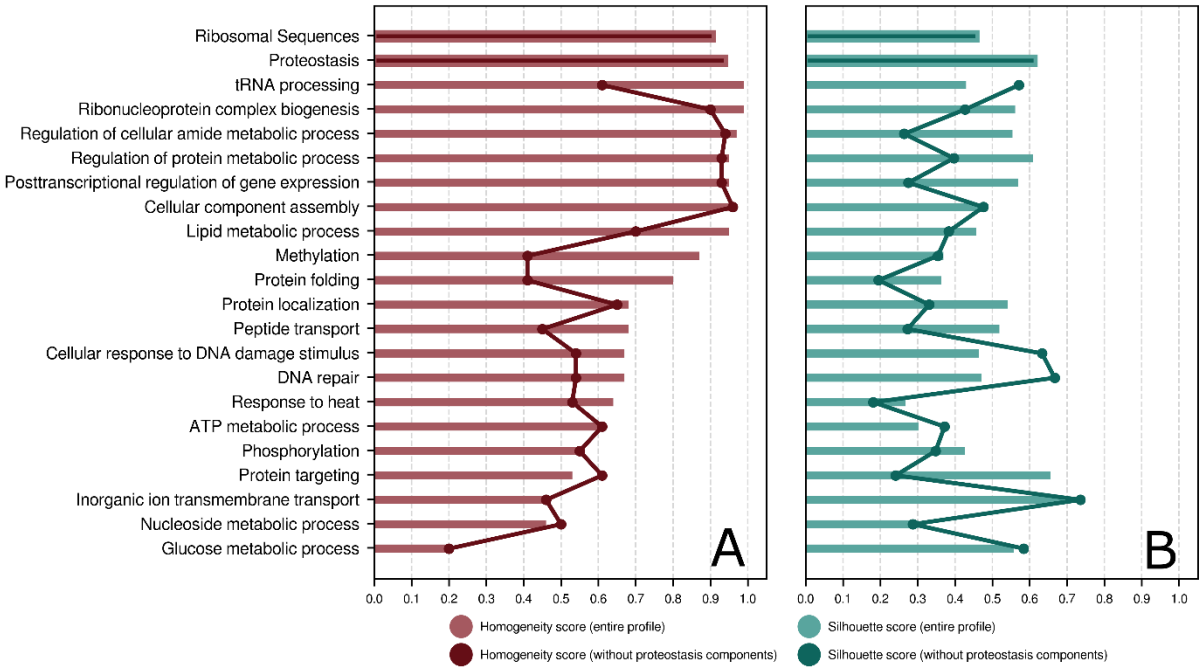
Figure 3



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

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

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)