Supporting information:
Physicochemical classification of organisms

ELOY VALLINA ESTRADA AND MIKAEL OLIVEBERG *

Department of Biochemistry and Biophysics, Arrhenius Laboratories of Natural Sciences, Stockholm University, S-106 91 Stockholm, Sweden.
*Corresponding author: mikael.oliveberg@dbb.su.se

Methods

Fig. S1. Data collection schemes from UniProt and NCBI. From each source, filters are applied and sequences collected, translated CDS in the case of Assemblies and protein sequences in the case of UniProt. In both cases, TaxIds are mapped to lineages and the whole genome GC content is read from NCBI’s genome reports. In either case, physicochemical features are calculated from the sequences.
Table S1. The protein features analysed in this study. Throughout the text, parameters referring to the average properties of a whole organism proteome are denoted by superscript ‘average’.

| property  | derivation | description |
|-----------|-------------|-------------|
| \( n_{\text{genes}} \) |  | Number of genes in the proteome. |
| \( l_{\text{sequence}} \) |  | Protein length expressed as number or residues. |
| MW | \( \sum_{i}^{20} n_{i} w_{i} \) | Protein molecular weight determined by the sum of the masses of its residues \( (w_{i}) \). |
| SASA | 6.3 MW\(^{0.73} \) | Solvent Accessible Surface Area estimated with Miller’s empirical expression. Units are in Å\(^2 \). |
| \( f_{\text{neg}} \) | \( \frac{n_{D} + n_{E}}{l_{\text{sequence}}} \) | Fraction of negatively charged residues, where \( n_{D} \) and \( n_{E} \) are the number of D and E, respectively. |
| \( f_{\text{pos}} \) | \( \frac{n_{R} + n_{K}}{l_{\text{sequence}}} \) | Fraction of positively charged residues, where \( n_{R} \) and \( n_{K} \) are the number of R and K, respectively. |
| \( f_{\text{charged}} \) | \( f_{\text{pos}} + f_{\text{neg}} \) | Fraction of charged residues of any sign. |
| \( f_{\text{hydrophobic}} \) | \( \frac{n_{F} + n_{L} + n_{I} + n_{V}}{l_{\text{sequence}}} \) | Fraction of hydrophobic residues, based on the numbers of F, L, I, and V. |
| \( \text{charge}_{\text{net}} \) | \( n_{K} + n_{R} - n_{D} - n_{E} \) | Difference between counts of positively and negatively charged residues. |
| NCD | \( \frac{\text{charge}_{\text{net}}}{\text{SASA}} \) | Amount of charge per square unit of protein solvent accessible surface. Units are in e/Å\(^2 \), where e represents the elemental electron charge. |
| GC |  | Percentage of guanine and cytosine of across an organism’s whole genome. |
Supplemental results and controls

Representing organisms in a space of physicochemical coordinates. Our strategy for comparing different organisms in terms of physicochemical properties is based on mapping each proteome to a specific set of coordinates, where each coordinate is the proteome-wide property average (Table S1). This is done under the assumption that the distribution of each property over the proteome is approximately unimodal, reasonably narrow, and symmetric. By analysing the cumulative density functions (CDF or the distribution integrals) of each variable for every proteome, we find that these assumptions hold remarkably well for NCD and \( f_{\text{charged}} \): most traces are smoothly sigmoidal (Fig. S2). The few proteomes that deviate from this trend tend to be small by containing just hundreds, rather than several thousands, of proteins.

**Fig. S2.** Cumulative density functions (CDF) of the physicochemical properties NCD, \( f_{\text{charged}} \), and MW for all protein members in each bacterial proteome (\( n = 13,652 \)). Each CDF is centred at the average value of the underlying data set, and this value is normalised to zero for easy comparison of shapes. Steeper transitions correspond to narrower distributions with smaller standard deviations. The few organisms with visible humps along the transitions indicate the presence of additional modes, i.e., their distribution deviates from unimodal. From these data, we conclude that and NCD and \( f_{\text{charged}} \) are distributed narrowly, symmetrically, and rather unimodally in most bacterial proteomes. The most significant asymmetries are observed for MW, where positive values account for about 70% of the observations, i.e., the distributions tail towards higher mass (c.f. manuscript Fig. 1).

However, symmetry breaks for MW, whose distribution is rather that of a power law and positively skewed (Fig. S2). Although a more robust measure of centrality would in this case
be the median, the differences between the median and the mean remain small (\(\sim 0.18\) standard deviations) even for proteomes with fewer than 500 proteins (\(\sim 0.24\) standard deviations). The result of employing the median as measure of centrality is shown in the NCD\textsubscript{median} - MW\textsubscript{median} plot (Fig. S3). Except for an overall shift of the plot towards lower MW, and a series of observations with a NCD\textsubscript{median} = 0, no significant changes occur (cf. main text Fig. 2).

![Fig. S3. Physicochemical map of life in terms of median NCD and MW. Although MW\textsubscript{median} is generally shifted down with respect to MW\textsubscript{average}, and some proteomes exhibit exactly NCD = 0, the relative positions of most observations are not changed (cf. main text Fig. 2).](image)

The role of histidines. Our estimates of protein net charge rely on simple counts of acidic (D and E) and basic residues (K and R), while we have disregarded the charge of histidines. With pK\textsubscript{A} values of 6.0 to 6.5, we would expect an average histidine charge of less than +0.5 at physiological pH, and this estimate is further complicated the histidine’s role in conjugating metals (1). Across all organisms, histidines are also among the least abundant residues: on average they appear twice per 100 residues, compared to 5 - 6 times for D, R, K and E. In other words, the histidines are expected to have a relatively small contribution to protein net charge. The effect of including the histidines in the analysis is shown in Fig. S4, where we have tentatively added +0.5 charge units per average number of histidines in each proteome. Although the scatter becomes shifted toward positive NCD\textsubscript{average} values (cf. main text Fig. 2), the relative positions of the majority of observations remain unchanged and the key features discussed in our study are still visible, underlining the robustness of the method.
Fig. S4. Net charge-density estimate with all histidines given a charge of +0.5. Since there are ~6 histidines per average-sized protein, i.e., 300 residues, the average net charge increases +3 units and the NCD\textsubscript{average} increases +0.17 (eÅ\textsuperscript{2} x 1000).

**Consistency between NCBI and UniProt data.** Analysis of the organisms present both in the NCBI and UniProt datasets allows us to evaluate the consistency of our results across the two sources. Comparison of the NCD\textsubscript{average}, MW\textsubscript{average} and \(f\)\textsubscript{charged}\textsubscript{average} values derived from the intersection of the two datasets shows good agreement (Fig. S5). Pronounced disparities appear only in the MW\textsubscript{average} of some eukaryotes, e.g., *Homo sapiens*, *Gallus gallus*, *Sus scrofa* and *Xenopus tropicalis*. The reason for the deviation is that these are richly annotated, popular model organisms, for which multiple isoforms of the same gene have been detected. Such isoforms, including the products of translational frameshifting (2) and alternative splicing (3), generate independent records in NCBI, whereas in UniProt they are typically grouped under a single entry. Thus, the more thoroughly a complex organism is investigated, the more the proteome composition will differ between the ‘all-inclusive’ NCBI and the ‘curated’ UniProt datasets.
Fig. S5. Mirror features over the intersection (n = 3,276) between the UniProt and NCBI datasets of proteomes. Off-diagonal points indicate disparities across the two representations of the same organism, the most conspicuous of which are labelled.

While the coordinates of organisms in the intersection are similar, a major difference between the two datasets lies in their size: there are three times more NCBI proteomes than UniProt proteomes. As a result, the UniProt proteomes have lower coverage of organisms, in particular when it comes to the organisms that remain poorly understood because they are difficult to grow and study under laboratory conditions. For instance, the functionally intriguing class *Ca. Poseidoniiia* that composes most of the archaenal Cluster 2 are not present in the taxonomic slice obtained from UniProt, save for a single member (c.f. main text Fig. 2 and Figs. S6-S8). Similarly, regarding the details of phylogenetic clustering, the reduced number of representatives in the phylum *Apicomplexa* makes their property diversion difficult to capture (Fig. S6).
Fig. S6. NCD\textit{average} vs. MW\textit{average} maps for \textit{Archaea}, \textit{Bacteria} and \textit{Eukaryota} derived from the UniProt Proteomes dataset (cf. main text Fig. 3). \textbf{A}. For \textit{Archaea}, the smaller size of the dataset ($n = 328$), leads to loss of two archaeal phyla and a radical reduction in the number of observations in the remaining classes of \textit{Euryarchaeota}. Most notably, this dataset does not include enough representatives of \textit{Ca. Poseidonii} to infer the existence of Cluster 2 (see Fig. 2 and Figs. S7-S8).
B. For Bacteria \((n = 8,056)\), the smaller UniProt dataset yields results very similar to those presented in the main text (Fig. 2), albeit that the bias towards negative NCD\textsubscript{average} is stronger \((84:16)\) than observed for the NCBI Assemblies \((76:24)\). c. Similar to Archaea, the UniProt sample of Eukaryota shows the same tendencies detected in the NCBI dataset despite the notable size reduction \((n = 1,609)\). Even so, the lower coverage precludes detection of the orderly divergence among Apicomplexa apparent in main text (Fig. 2).

**Archaea splits into three clusters.** The archaeal dataset shows clustering in the NCD\textsubscript{average} - MW\textsubscript{average} plot, where the *Halobacteria* in Cluster 1 most clearly stands out (main text Fig. 3). To better separate the two clusters in central area of the plot, we treated the data as a mixture of three two-dimensional Gaussian distributions, assigning each observation to the subpopulation from which it most probably originates (4). Because of the scale difference between the two dimensions, the data was first rescaled to the ranges of \((-1, 1)\) and \((0, 1)\) for NCD\textsubscript{average} and MW\textsubscript{average}. The result of passing back the cluster labels to the observations in the original coordinates is shown in Fig. S7.

![Archaea](image)

**Fig. S7.** The archaeal subset of NCBI Assemblies splits into three clusters by application of a Gaussian Mixture Model (GMM) (cf. main text Figs. 2 and 3). Under the assumption of three subpopulations in the NCD\textsubscript{average} - MW\textsubscript{average} plot, GMM determines to which subpopulation each observation most likely belongs. The box marks 5 outliers, where the two crosses indicate halobacterial assemblies (for further information, see section ‘A closer look at outliers’ below).
Another strategy for splitting the archaea data into clusters involves Principal Component Analysis (PCA), performed here on 9 physicochemical properties $MW_{\text{average}}$, $NCD_{\text{average}}$, $f_{\text{pos}}^{\text{average}}$, $f_{\text{neg}}^{\text{average}}$, $f_{\text{charged}}^{\text{average}}$, $f_{\text{hydrophobic}}^{\text{average}}$, $\text{charge}_{\text{net}}^{\text{average}}$, $n_{\text{genes}}^{\text{average}}$ and GC (Table S1). Essentially, PCA is a dimensionality reduction technique that allows projection of the 9-dimensional space onto a two-dimensional plane with preservation of the separating features. In the PCA projection (Fig. S8), Cluster 1 Nanohaloarchaea separates from Halobacteria due to their differences in GC and $f_{\text{charged}}^{\text{average}}$ content (cf. main text Fig. 6), and the separation between Ca. Poseidonii in Cluster 2 and the species in Cluster 3 becomes overall clearer (cf. Fig. S7 and main text Fig. 3a).

![Euryarchaeota](image)

**Fig. S8.** The first two principal components PC1 and PC2 of the archaeal dataset. The highlight shows the main taxonomical classes of phylum *Euryarchaeota*.

**Bacterial endosymbionts are enriched in nucleic acid binding proteins.** The proteomes of bacteria with conspicuously positive $NCD_{\text{average}}$ reveal a common lifestyle. Of 133 bacteria scoring $NCD_{\text{average}} > 1 \times 10^{-3}$ eÅ$^{-2}$, 122 belong to different genera of insect endosymbionts, namely *Buchnera* (32), *Blattabacterium* (21), *Ca. Hodgkinia* (15), *Ca. Sulcia* (14), *Ca. Portiera* (12), *Ca. Carsonella* (9), *Ca. Tremblaya* (8), *Ca. Nasuia* (3) and *Ca. Nardonella* (3). These endosymbionts carry remarkably small proteomes, containing only $\sim 1/10$ of the bacterial average of 3,612 proteins. Such proteome reduction is consistent with their mutualistic lifestyle and leaves mainly genes for the central processes of replication,
transcription, translation, and cell division (5, 6). To shed light on the bias towards positively charged proteins, we compared the protein identities of 10 of the endosymbionts listed above with 10 free-living bacteria. In short, each protein in the organisms was mapped to its Gene Ontology (GO) terms and we counted the number of proteins listing at least one GO-term connected to ‘nucleic acid binding’ function: GO-0003676 or any of its children. The results suggest that bacterial endosymbiotic proteomes are, on average, twice as rich in proteins that bind RNA and DNA as free-living bacteria (Fig. S9). Underlining the functional importance of these proteins, a negative correlation between proteome size and fraction of nucleic acid binding proteins is seen for both groups (Fig. S9).

**Fig. S9.** Endosymbiotic bacteria are characterized by small proteomes enriched in positively charged nucleic binding proteins. A. Fraction of nucleic acid binding proteins (GO-0003676 and children) for representative organisms. B. Correlation between proteome size and fraction of nucleic acid binding proteins therein: the smaller the proteome, the higher the fraction of nucleic acid binding proteins.

**Physicochemical differences at lower taxonomic levels in Eukaryota.** Representative separations at class level are observed within the eukaryotic phyla Nematoda (roundworms) and Basidiomycota (club fungi) (Fig. S10). In the former phylum, the classes Enoplea and Chromadorea separate mainly due to differences in NCDAverage. The class Enoplea contains, for example, the pathogenic parasite Trichinella, which is responsible for the disease trichinosis, associated with the ingestion of undercooked meats (7). Within Chromadorea, we find the important model organism Caenorhabditis elegans. In the phyla Basidiomycota,
the class *Agaricomycetes* (mushroom-forming fungi) stands out due to $\text{MW}^{\text{average}}$ values that are lower than those for *Tremellomycetes* (jelly fungi), *Wallemiomycetes* (xerophilic mould) and *Microbotryomycetes* (yeast and plant pathogens) (Fig. S10).

**Fig. S10.** Separations between taxonomic classes within *Nematoda* (roundworms) and *Basidiomycota* (club fungi). A. The two nematode classes *Chromadorea* and *Enoplea* are mainly separated by differences in $\text{NCD}^{\text{average}}$. B. The class *Agaricomycetes* separates from other *Basidiomycota* due to its relatively small proteins, i.e., $\text{MW}^{\text{average}}$.

**Physicochemical differences at lower taxonomic levels in Bacteria revealed by the alternative parameters $f^{\text{average}}_{\text{charged}}$ and GC.** Additional resolution of physicochemical preferences is obtained by plotting bacteria on the $f^{\text{average}}_{\text{charged}}$-GC plane (Fig. S11, cf. main text Fig. 4). Although some correlation between the two properties is apparent, the bacterial clusters remain clearly separated. At the level of the phyla, the numerous *Proteobacteria* span the whole range of GC values, while *Firmicutes*, *Actinobacteria*, *Bacteroidetes* and *Tenericutes* are constrained to particular intervals. By dividing *Proteobacteria* into classes, the picture gains even further detail: $\gamma$-proteobacteria are biased to relatively low $f^{\text{average}}_{\text{charged}}$ across a wide range of GC values, and $\delta$- and $\varepsilon$-proteobacteria show higher $f^{\text{average}}_{\text{charged}}$ with more narrow and well separated spans of GC values (Fig. S11).
Fig. S11. GC - $f_{\text{charged}}$ average plots of Bacteria. At the level of phyla, some stacking in terms of GC is observed for Actinobacteria, Firmicutes and Tenericutes, while Proteobacteria and its classes span the whole range of GC values.

**Link between proteome properties and organism growth temperature.** An organism’s ability to sustain extremely high temperatures can be seen to correlate with preferences at the level of some proteome physicochemical features. In particular, organisms growing at temperatures higher than 60 °C show a tendency towards high $f_{\text{charged}}$ average and low GC (Fig. S12 cf. main text Figure 5). In this sense, the model organism *T. thermophilus* constitutes an exception rather than the rule, which underlines the importance of extensive sampling when exploring niche adoptions. Conversely, the psychrophiles ($T_{\text{growth}} < 15$ °C) shows $f_{\text{charged}}$ average values that seem relatively low, albeit that the few species in this group makes the statistics uncertain (Fig. S12 cf. main text Figure 5).
Fig. S12. Variation of physicochemical properties with growth temperature ($T_{\text{growth}}$) for thermophilic ($n = 45$), mesophilic ($n = 7,722$), and psychrophilic ($n = 13$) bacteria. A. Thermophiles are distinguished by their high $f_{\text{charged}}$ average, where the model organism *T. thermophilus* ($T_{\text{growth}} = 75^\circ\text{C}$ and $n = 6$) shows a relatively low value of 0.24. B. Conversely, the thermophiles display relatively low GC, where *T. thermophilus* is an outlier with a value of 70%. The remaining two observations in the outlying cluster are *Thermus parvatiensis* ($T_{\text{growth}} = 80^\circ\text{C}$, GC = 69%) and *Marinithermus hydrothermalis* ($T_{\text{growth}} = 70^\circ\text{C}$, GC = 69%).

Additional examples of UMAP separations and backtracking of the underlying parameters. The benefit of applying UMAP to our datasets is best illustrated by examining the cases where the separations can be directly traced back to the NCD average–MW average and GC–$f_{\text{charged}}$ average projections. In *Archaea*, for instance, UMAP reveals clearer separation of *Euryarchaeota* at class level: *Halobacteria* and *Nanohaloarchaea* in Cluster 1, *Ca. Posedoinia* in Cluster 2, as well as *Thermococci* and *Methanomicrobia* in Cluster 3 are all better segregated (Fig. S13 cf. Fig. S8 and main text Fig. 3). Corresponding results are obtained at the level of orders, as illustrated for *Methanomicrobia* in the crowded Cluster 3 (Fig. S13). Although the result shows that differences between closely related organisms are readily distinguished, a shortcoming of the UMAP projection is that it conceals in which physicochemical property the distinction lies. Nevertheless, the UMAP separations can in this case be traced back to linear plots of the base parameters NCD average, MW average, GC and $f_{\text{charged}}$ average for determination of their respective influence (c.f. main text Fig. 3).
Fig. S13. Classes within the archaeal phylum *Euryarchaeota*, where the class *Methanomicrobia* is further separated into orders. **A.** The UMAP projection separates clearly *Halobacteria* from *Nanohaloarchaea*, and *Ca. Poseidonila* from other archaea. In the right panel, the class *Methanomicrobia* is split into taxonomical orders. **B-C.** Back-tracing of data in panel A. The UMAP separations between the orders *Methanosarcinales* and *Methanomicrobiales* are mainly traced back to differences in GC and $f_{\text{charged}}$.

A full series of bacterial UMAP separations is shown in main text Fig. 6, starting at the phylum *Proteobacteria*, and spanning the five taxonomic levels *phylum, class, order, family* and *genus*. In this case, *Klebsiella* are distinguished from sibling genera of the family *Enterobacteriaceae* due to relatively high GC (Fig. S14).
Fig. S14. Proteome-property differences between the bacterial genus *Klebsiella* and its sibling genera. **A.** All genus exhibit NCD$^\text{average}$ around zero and similarly large spread in MW$^\text{average}$ (c.f. the extended UMAP clusters in main text Fig. 6). **B.** Corresponding differences in GC and $f^\text{average}_\text{charged}$, where *Klebsiella* stands out at the high-GC end.

In the same way, UMAP accentuates the distinction of the obligate bacterial endosymbionts (5, 6). While these organisms are phylogenetically diverse and spread over different MW$^\text{average}$, GC and $f^\text{average}_\text{charged}$ values, they cluster peripherally in the UMAP plane mainly due to their positive NCD$^\text{average}$ (Fig. S14).

Fig. S15. UMAP projection of bacterial endosymbionts and back tracing of their physicochemical properties. **A.** The different genera of bacterial endosymbionts are segregated into a tight cluster in the UMAP representation. **B.** Preferred values of $f^\text{average}_\text{charged}$ and GC are consistent within genera and markedly different across the taxa. **C.** All of these endosymbionts have a stark positive bias in terms of NCD$^\text{average}$, in contrast to most other bacteria.
Our final UMAP example concerns the separation of the many kinds of fungi in the eukaryote dataset. Most strikingly, the budding yeast (*Saccharomyces*) appears clearly differentiated from all the other sac fungi in the phylum *Ascomycota* (Fig. S15). Moreover, the four major classes within the phylum *Basidiomycota* (club fungi) fall orderly in different regions of the plot, where *Wallemiomycetes* seems to split between two distinct clusters (Fig. S16). The separation of these two clusters is linked to speciation within the genus *Wallemia* (8): the left cluster contains 25 different entries of *W. mellicola*, a food contaminant with pathological effects in humans (9), and the right contains 21 different entries of *W. ichthyophaga* (halophilic fungus) (10), as well as a single observation of its close sibling *W. hederae* (11). With respect to their proteome properties, the three species differ mainly in GC and MW average (Fig. S17).

**Fig. S16.** UMAP separations of fungi classes and species within Eukaryota. A. *Ascomycota* splits into two groups, one of which matches exclusively the class *Saccharomyces* (budding yeasts). B. Within club fungi (*Basidiomycota*), the class *Wallemiomycetes* is made of a single order, *Wallemiales*; a single family, *Wallemiaceae*; and a single genus *Wallemia*. Representatives of the two main species of *Wallemia* form distinct clusters in the UMAP embedding. For backtracking of *Wallemia* protein properties, see Fig. S17.
Species of the eukaryotic genus *Wallemia* mapped to the physicochemical properties of their proteomes. **A.** Plot of NCD\textsuperscript{average} vs. MW\textsuperscript{average}. **B.** Plot of \( f_{\text{charged}} \) vs. GC. The multiple observations for the three species *W. mellicola* \((n = 25)\), *W. ichthyophaga* \((n = 21)\), and *W. hederae* \((n = 1)\) have so similar properties that they superimpose into single points. This single-point appearance contrasts the corresponding UMAP plots (Fig. S18), where the minimum-distance parameter of 0.99 separates the superimposed points into clusters.

A closer look at outliers. Although most organisms show consistent clustering in the physicochemical landscape, there are some notable outliers that break the patterns. In *Archaea*, five observations are highlighted in Fig. S7 due to their low MW\textsuperscript{average} < 15 kDa. Two of these outliers are annotated as *Halobacteria* (GCA_003226235 and GCA_006864335) but deviate from the other members of this taxonomic class by having slightly positive NCD\textsuperscript{average} (cf. main text Fig. 3). Upon examination, both assemblies indicate questionable quality as they contain large numbers of proteins derived from pseudogenes (79 % and 82 %, respectively). Moreover, these assemblies are excluded from RefSeq because of their high abundance of frameshifted protein and NCBI’s taxonomy check turned out inconclusive. The other three highlighted assemblies correspond to two unclassified archaea (GCA_004145285 and GCA_004173275) and one *Thaumarchaeota* (GCA_002763195), all of which show quality deficiencies such as fragmentation and inconclusive taxonomy. Another deviating organism is *Haloquadratum walsbyi* in the archaeal Cluster 1, where the five deposited assemblies show markedly low GC (main text Fig. 4). In this case, however, the assembly quality of *Haloquadratum walsbyi* is high, and the physiological features and morphology of this organism are known to be distinct from the other members of *Halobacteria* (12).
In *Bacteria*, the NCD\textsubscript{average}-MW\textsubscript{average} plot contains some observations at MW\textsubscript{average} < 16 kDa that match different species of *Mycoplasma*, i.e., *M. putrefaciens*, *M. edwardii*, *M. synoviae*, and *M. alkalescens* (Fig. S18). All of these have kindred representatives with more common MW\textsubscript{average} between 32 and 48 kDa (Fig. S18). Although the assemblies themselves do not show signs of bad quality, they contain twice as many proteins (~1400) as the other members of the genus (~700), many of which are labelled ‘uncharacterised’. In combination with the small protein size, this doubling in \( n_{\text{genes}}^{\text{average}} \) suggests that the additional proteins are very small, pointing either at problems in the construction of the assemblies or at artefactual translation of non-coding open reading frames. The most remarkable outliers in *Bacteria* are nonetheless the endosymbionts (Fig. S15). Besides being distinct by their positive NCD\textsubscript{average} values, the genus *Ca. Carsonella*, *Ca. Nardonella* and *Ca. Nasuia* show GC below 20% (Fig. S15).

![Figure S18](image-url)  

**Fig. S18.** In *Bacteria*, some observations stand out because of their low protein size. **A.** Several of the assemblies with low MW\textsubscript{average} correspond to species of *Mycoplasma*, with more common siblings at substantially higher MW\textsubscript{average} and similar NCD\textsubscript{average}. **B.** Both high- and low-MW\textsubscript{average} *Mycoplasma* exhibit diverse and orderly GC and \( f_{\text{charged}}^{\text{average}} \) patterns, similar to the majority of other *Bacteria*.

**Horizontal gene transfer and physicochemical analysis of an individual type of proteins: ribonucleotide reductase (RNR).** Ribonucleotide reductases catalyse the formation of deoxyribonucleotides from ribonucleotides, a key first step in the synthesis of DNA that is conserved across virtually all cellular organisms (13). The RNRs provide thus a good system for illustrating how diverging proteome properties manifest at the level of individual proteins. Phylogenetic studies suggest, moreover, that the RNRS in *Halobacteria* originate from
horizontal gene transfer between *Salinibacter ruber* and several possible archaeal acceptors, e.i. *Natronomonas pharaonis, Halorubrum lacusprofundi, Halomicrobium mukohataei* (13). From a physicochemical standpoint, *S. ruber* constitutes an ideal donor for this horizontal gene transfer to take place: not only does it share the ecological niche of *Halobacteria*, but also it exhibits a NCDaverage value at the extreme negative end of the bacterial distribution, i.e., at minimal physicochemical distance from *Halobacteria* (Fig. S19). The picture is further clarified by comparing specifically the structure of the large subunit of human RNR1 (UniProt: P23921, PDB: 3hnc), with those of *Escherichia coli* (UniProt: P00452, PDB: 2xap) and *Halobacterium salinarum* (UniProt: Q9HMU3, homology modelled on PDB 5im3—*Pseudomonas aeruginosa*). The results show that the surface properties of the individual proteins follow those of the proteomes, albeit with a varying offset to higher $f_{pos}$ and $f_{neg}$ values (Fig. S19).

![Fig. S19](image-url)  
**Fig. S19.** Physicochemical matching in the context of RNR horizontal gene transfer (HGT).  
**A.** The bacterial donor (*Salinibacter ruber*, two assemblies) for horizontal gene transfer exhibits extreme NCD average properties, similar to those of the *Halobacteria* acceptors (*Natronomonas pharaonis, Halorubrum lacusprofundi, Halomicrobium mukohataei*) (13).  
**B.** Comparison of the fractional abundance of positively (K and R) and negatively charged (D and E) residues for the individual RNRs and their whole proteomes. Circles denote proteomes, crosses RNRs, and the dotted line denotes net charge = 0. The charge of the RNRs follow the average of their proteomes, albeit with a varying offset to higher $f_{pos}$ and $f_{neg}$ values.  
**C.** Differences in surface electric surface potential of the RNR molecules as derived from their crystal structures: *H. salinarum* (UniProt Q9HMU3 modelled on PDB 5im3), *S. ruber* (UniProt Q2S5T5 modelled on 1xjk), *E. coli* (PDB 2xap), and *H. sapiens* (PDB 3hnc).
**Amino acid preferences in halophilic archaea.** Within *Archaea*, *Halobacteria* stand out due to their remarkably negative NCD\(^{\text{average}}\). PCA of the amino acid fractional abundances allows us to determine which changes to protein surface composition have taken place during the evolution of these organisms. Fig. S20 shows that the main determinants of the high-charge adaptation in *Halobacteria* are enrichment in aspartate (D) and depletion of lysine (K). In contrast, the selective pressure against K seems weaker in the sibling taxon *Nanohaloarchaea*, where the low NCD\(^{\text{average}}\) values stem mainly from preferential enrichment of glutamate (E) (Fig. S20). The observation is also reflected by the increased \(f_{\text{charged}}^{\text{average}}\) of *Nanohaloarchaea*, shown in main text Fig. 4.

**Fig. S20.** Preferences for charged amino acids in halophilic *Archaea*. A. Principal component analysis (PCA) of the average amino-acid composition of *Archaea*. Blue arrows indicate the orientation of the feature vectors with respect to the PC1-PC2 plane, where observations farther out show enrichment in the corresponding amino acids. B. Average abundance of charged amino acids in the different groups of halophilic *Archaea*.

**Coupling between fraction of hydrophobic residues and NCD\(^{\text{average}}\).** In addition to charge, hydrophobic residues exposed to the solvent are expected to play a crucial role in modulating protein-protein interactions. Since the majority of hydrophobic side chains appear in the proteins’ core, estimates of their surface exposure from sequence data alone are more uncertain than corresponding estimates of surface charge. For example, \(f_{\text{hydrophobic}}^{\text{average}}\) is expected to have a strong basal dependence on protein size as the volume scales faster than the surface. To minimise the impact of this geometrical skewing, we present here data for
Archaea where the span of $MW^{\text{average}}$ values is relatively small. The result can be taken to indicate that organisms mitigate intracellular protein aggregation by increasing repulsive charge and decreasing surface hydrophobicity concomitantly. Taking the interpretation one step further, the curvature caused by the Halobacteria offset may even reflect the regain of electrostatic repulsion in the ionic-liquid regime, i.e., the predicted screening-length increase in Halobacteria (14). The data shows also that $f_{\text{hydrophobic}}^{\text{average}}$ follows $f_{\text{positive}}^{\text{average}}$, whereas it anti-correlates with $f_{\text{negative}}^{\text{average}}$, shedding light on how $NCD^{\text{average}}$ is regulated (SI Movie 6).

Although these observations are intriguing, they are currently left out from detailed discussion because they rely on overly simplified estimates where the systematic errors cannot satisfactorily be assessed.

![Graph](image)

**Fig. S21.** The relation between $f_{\text{hydrophobic}}^{\text{average}}$ and $NCD^{\text{average}}$ exemplified for Archaea. Decrease in $f_{\text{hydrophobic}}^{\text{average}}$ accompanies accumulation of negative charge in protein surfaces for most Archaea, while the effect caps off below $-1.5 \text{ eÅ}^{-2}$, in the halophilic region.
References

1. Arnold FH & Haymore BL (1991) Engineered metal-binding proteins: purification to protein folding. *Science* 252(5014):1796.
2. Atkins JF, Loughran G, Bhatt PR, Firth AE, & Baranov PV (2016) Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Research* 44(15):7007-7078.
3. Pan Q, Shai O, Lee LJ, Frey BJ, & Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40(12):1413-1415.
4. Duda RO & Hart PE (1973) *Pattern classification and scene analysis* (Wiley New York).
5. Bennett GM, McCutcheon JP, MacDonald BR, Romanovicz D, & Moran NA (2014) Differential Genome Evolution Between Companion Symbionts in an Insect-Bacterial Symbiosis. *mBio* 5(5):e01697-01614.
6. McCutcheon JP & Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10(1):13-26.
7. Bruschi F & Murrell KD (2002) New aspects of human trichinellosis: the impact of new &lt;em&gt;Trichinella&lt;/em&gt; species. *Postgraduate Medical Journal* 78(915):15.
8. Zalar P, Sybren de Hoog G, Schroers HJ, Frank JM, & Gunde-Cimerman N (2005) Taxonomy and phylogeny of the xerophilic genus Wallemia (Wallemiomycetes and Wallemiales, cl. et ord. nov.). *Antonie Van Leeuwenhoek* 87(4):311-328.
9. Zajc J & Gunde-Cimerman N (2018) The Genus Wallemia-From Contamination of Food to Health Threat. *Microorganisms* 6(2).
10. Zajc J, Kogej T, Galinski EA, Ramos J, & Gunde-Cimerman N (2014) Osmoadaptation strategy of the most halophilic fungus, Wallemia ichthyophaga, growing optimally at salinities above 15% NaCl. *Appl Environ Microbiol* 80(1):247-256.
11. Janečič S, et al. (2016) Halophily reloaded: new insights into the extremophilic lifestyle of Wallemia with the description of Wallemia hederae sp. nov. *Fungal Diversity* 76(1):97-118.
12. Bolhuis H, et al. (2006) The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. *BMC Genomics* 7(1):169.
13. Lundin D, Gribaldo S, Torrents E, Sjöberg B-M, & Poole AM (2010) Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. *BMC Evolutionary Biology* 10(1):383.
14. Wennerstrom H, Vallina Estrada E, Danielsson J, & Oliveberg M (2020) Colloidal stability of the living cell. *Proc Natl Acad Sci U S A* 117(19):10113-10121.