Census of exposed aggregation-prone regions in proteomes

Théo Falgarone, Étienne Villain, Francois Richard, Zarifa Osmanli, Andrey V Kajava

Census of exposed aggregation-prone regions in proteomes

Théo Falgarone[1], Étienne Villain[1], Francois Richard[1], Zarifa Osmanli[1,2] and Andrey V. Kajava[1,3]*

[1]Centre de Recherche en Biologie cellulaire de Montpellier, CNRS, Université Montpellier, Montpellier, 34293, France, [2]Biophysics Institute, Ministry of Science and Education of Azerbaidjan Republic, Az1141, Baku, Azerbaidjan, [3]Institut de Biologie Computationnelle, Université Montpellier, 34095 Montpellier, France

* To whom correspondence should be addressed. Email: andrey.kajava@crbm.cnrs.fr

Keywords: protein aggregation, bioinformatics, large-scale analysis, evolution, kingdoms of life.

ABSTRACT

Loss of solubility usually leads to the detrimental elimination of protein function. In some cases, the protein aggregation is also required for beneficial functions. Given the duality of this phenomenon, it remains a fundamental question how natural selection controls the aggregation. The exponential growth of genomic sequence data and recent progress with *in silico* predictors of the aggregation allows approaching this problem by a large-scale bioinformatics analysis. Most of the aggregation-prone regions are hidden within the 3D structures and, therefore, they cannot realize their potential to aggregate. Thus, the most realistic census of the aggregation prone regions requires crossing aggregation prediction with information about the location of the natively unfolded regions. This allows us to detect so-called "Exposed Aggregation-prone Regions" (EARs). Here, we analyzed the occurrence and distribution of the EARs in 76 full reference proteomes from the three kingdoms of life. For this purpose, we used the TAPASS pipeline, which provides a consensual result based on several predictors of aggregation. Our analysis revealed a number of new statistically significant correlations about the presence of EARs in different organisms, their dependence on protein length, cellular localizations, co-occurrence with short linear motifs, and the level of protein expression. We also obtained a list of proteins with the conserved aggregation-prone sequences for further experimental tests. Insights gained from this work led to a deeper understanding of the functional and evolutionary relations of the protein aggregation.

INTRODUCTION

Proteins are usually soluble molecules interacting transiently with each other or the other biomolecules. After performing their functions, they are degraded by proteases. Thanks to the dynamic balance between protein synthesis and degradation, living organisms can efficiently regulate many different processes. However, occasionally, some proteins, often for not entirely clear reasons, form aggregates. Most of the aggregates have a very characteristic structure of amyloid fibrils. These fibrils are typically straight, around 10 nm in diameter, thermostable, protease resistant, and rich in β-structure (1). They are completely or partially insoluble and frequently lead to a variety of age-related diseases including Alzheimer's disease, Parkinson's disease and others (2). In some cases, the amyloid fibrils (named prions) can be "infectious agents". The prion fibrils, which are found themselves in another organism or a cell, can trigger the formation of similar fibrils and cause transmissible neurodegenerative diseases (3). The amyloid deposits can not only be composed of copies of the same protein, but also represent co-aggregates of two or more proteins and by doing so simultaneously impair several biological processes(4). At the same time, not all amyloid fibrils are linked to diseases. Increasing number of studies describe so called "functional" amyloids, which fulfill beneficial roles in the organism (5, 6). For example, curli proteins from some gram-negative bacteria

form amyloid fibrils on the bacterial surface. They are involved in biofilm formations, which is a successful strategy allowing microorganisms to resist the threats of the environment (UV radiation, oxygen, desiccation etc) (7). Other examples from mammals are RIP1 and RIP3 proteins whose co-aggregation into amyloid fibrils mediates a key interaction of necroptosis signaling (8, 9).

Despite great interest in protein aggregation, especially regarding amyloids, scientists have focused on a few of the most devastating amyloidoses or known cases of functional amyloids. However, the overall prevalence of the protein aggregation in organisms is not yet well studied. This analysis requires computational methods for *in silico* prediction of the aggregation. The propensity to form aggregates is coded by the amino acid sequence, therefore, several computational programs have been developed (10–18) Availability of the computational tools for prediction of aggregation-prone regions made it possible to obtain a more general view of this phenomenon by using *in silico* analysis of the whole-proteome data. Previous *in silico* studies revealed a number of interesting observations (14, 19–28). For example, a study of six proteomes (*P. tetraurelia, S. cerevisiae, C. elegans, D. melanogaster, M. musculus* and *H. sapiens*) using a specially developed algorithm, demonstrated that the average aggregation propensity of a proteome correlates inversely with the complexity and longevity of the studied organisms (29). In another analysis of the proteomes of *D. melanogaster*, *S.cerevisiae* and *C. elegans* using TANGO predictor(13) it was shown that proteins that are essential to organism fitness (knockdown of these genes leads to lethality), have a lower aggregation score than nonessential proteins(23). Analysis of the human proteome by the Zyggregator method(30, 31) suggested that proteins involved in the secretion pathway are more prone to aggregation compared to non-membrane proteins in general(27). Application of the 3D profile method to *E. coli*, *S. cerevisiae*, and *H. sapiens* proteomes showed that the predicted high propensity for amyloid formation does not reflect well the limited number of proteins involved in disease-related or functional amyloid deposits (26). The same analysis of proteins from PDB suggested that most of the predicted aggregation prone regions are hidden within the 3D protein structure and, therefore, inaccessible for intermolecular interactions such as aggregation (32). The analysis of cytosolic bacterial (*E. coli*) and eukaryotic (*H. sapiens*) proteomes indicated that the aggregation propensity of proteins inversely-correlates with their abundance (19–21). Most of these data are in agreement with the conclusion that the evolutionary pressure acts on the proteins to minimize their aggregation propensity.

Several publications have reported that proteomes contain a very high percentage of proteins with amyloidogenic or aggregation-prone regions (AR), which is in obvious conflict with a small number of the known proteins involved in amyloidoses(28). It was explained by the fact that most of the predicted ARs are hidden within the 3D structure preventing aggregation (33) . Conversely, in most known cases of amyloidosis, the native conformation of the polypeptide chains that form amyloid deposits *in vivo*, is unfolded (or intrinsically disordered). Thus, to get a more realistic census of the aggregation prone regions in proteomes, it is necessary to cross aggregation prediction with information about the location of the intrinsically disordered regions (IDRs). IDRs are always exposed for the intermolecular interactions critical for aggregation. We used this concept to develop a computational pipeline TAPASS (33), which can detect such “Exposed Amyloidogenic Regions” or, otherwise “Exposed Aggregation-prone Regions” (EARs) located within IDRs and carrying high potential to aggregate (see Figure 1). To obtain the most consensual results on the occurrence and distribution of the EARs in proteomes we selected three predictors of aggregation (TANGO, Pasta 2.0 and ArchCandy 2.0) (10, 13, 16). They were selected based on the diversity of their basic principles, their popularity, and ability to be downloaded for analysis of a large number of sequences. TAPASS also provides information about the cellular localization, post-translational modifications, and functions of aggregation-prone proteins.

In addition to the advances with the predictors of aggregation, the past few years were marked by a significant increase in the number and quality of proteome sequencing data, Thus, the advances with methods predicting aggregation potential, development of the TAPASS pipeline, as well as an increasing

number of high quality whole-proteome sequencing data, made a new census of aggregation-prone regions in proteins timely. In this paper, we present the results of such a detailed analysis of 76 full reference proteomes from the UniProt databank.

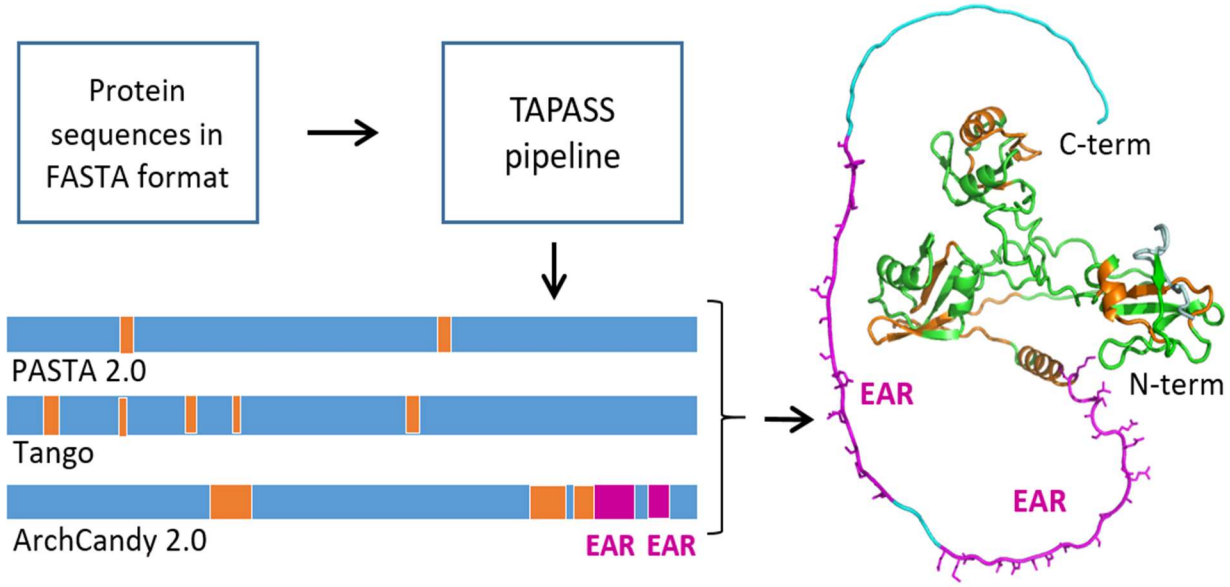


Figure 1. *A general scheme showing mapping of ARs and EARs on a structural model of human TAR DNA-binding protein 43. This protein forms amyloid fibrils by the C-terminal Low Complexity Domain (LCD, 274–414) (34). TAPASS predicts several ARs, which are located within the 3D structures (orange) and two EARs (magenta) located at the C-terminal IDR.*

MATERIAL AND METHODS

TAPASS pipeline

The input file of TAPASS requires protein sequences in Fasta format and can contain additional informations from UniProt(35) (gene id, GO term, version, modification date...). The pipeline uses: IUPred(36) and our in-house predictor (IDRs), CATH associated with HMMER 3.3 (structural domains)(37, 38), TMHMM (transmembrane regions)(39), SignalP (signal peptide)(40), SLiMs (short linear motifs)(41, 42), Pfam (structural and functional domains)(43), Pasta 2.0(16), TANGO(13) and updated version of ArchCandy 2.0 (aggregation-prone regions)(10, 33).

The results of the three predictors of aggregation, ArchCandy 2.0, Pasta 2.0 and TANGO, were treated separately. Each predictor gives the start and end positions of ARs in protein sequences. An AR is considered as EAR if at least 80 % of an individual hit of AR predictor overlap with an IDR. Thus, our analysis led to three independent censuses of the aggregation-prone regions. If all three censuses yielded similar regularities, then these findings were considered as more reliable and treated with special attention.

Selection of proteomes for large-scale analysis

76 reference proteomes with 1 123 749 proteins in total were selected from the UniProt databank (see Supplementary data) (35). The proteomes belong to the three kingdoms of life: eukaryote, bacteria and archaea. The selection of species was made to have well-annotated and complete reference proteomes covering the diversity of living organisms. Viral proteomes were not considered in this analysis due to small size of their proteomes yielding very different results depending on the strains. Their analysis will be a

subject of our future study.

RESULTS

Occurrence of ARs and EARs in the proteomes

Previous studies detected a very high percentage of AR-containing proteins in proteomes with almost each protein having at least one predicted AR (22, 26, 27). The results of our analysis of 76 reference proteomes support this conclusion predicting 68.6 %, 79.3 % and 90.0 % of AR-containing proteins by ArchCandy 2.0, Pasta 2.0 and TANGO, respectively. The coverage of ARs, obtained by dividing the number of amino acid residues involved in ARs by the number of all residues in proteins, is equal to 12.6 %, 6.2 % and 11.3 % for ArchCandy 2.0, Pasta 2.0 and TANGO respectively. A very high percentage of AR-containing proteins is in contradiction with a small number of proteins known to be involved in different amyloidoses or functional amyloids. However, if we consider EARs, the number of potential aggregation-prone proteins is drastically reduced. EAR-containing proteins represent 9.0 %, 6.8 % and 19.5 % of all proteins with coverage of 0.8 %, 0.2 % and 0.4 % of residues according to ArchCandy 2.0, Pasta 2.0 and TANGO respectively. The low percentage of proteins with EARs, in contrast to a very high percentage of ARs, agrees better with the small number of the known proteins involved in aggregation *in vivo*.

Aggregation-prone regions in prokaryotic and eukaryotic organisms

Analyzing the 76 selected proteomes we observed a relatively uniform distribution of AR-containing proteins among the organisms (Figure 2). Curiously, *Homo sapiens* has the least number of AR-containing proteins. At the same time, we saw a large variation in the proportion of EAR-containing proteins. Among the organisms with the least number of EAR-containing proteins are thermophilic prokaryotes (6 archaea and 5 bacteria: *Chloroflexus aurantiacus, Thermodesulfovibrio yellowstonii, Dictyoglomus turgidum, Nanoarchaeum equitans, Sulfolobus solfataricus, Thermotoga maritima, Archaeoglobus fulgidus, Thermococcus kodakaraensis, Methanocaldococcus jannaschii, Candidatus korarchaeum, Aquifex aeolicus*).

The eukaryotes with the simplest level of organization, mostly unicellular (or partially unicellular) protists such as *Plasmodium falciparum, Leishmania major, Thalassiosira pseudonana, Trypanosoma cruzi, Toxoplasma gondii and Dictyostelium discoideum* have the greatest numbers of EAR-containing proteins (Figure 2). High levels of EAR-containing proteins are also found in two fungi (*Ustilago maydis, Neurospora crassa),* fruit flies (*Drosophila melanogaster),* mosquitoes (*Anopheles gambiae*) and chickens (*Gallus gallus*). Most of them are known to have the greatest number of low-complexity repetitive sequences (44). This is particularly the case of *Trypanosoma cruzi* and *Dictyostelium discoideum*, which have an abnormal high level of Asn/Gln rich regions, two types of amino acids frequently found in amyloids. Among analyzed mammalians, *Homo sapiens* has the least number of EAR-containing proteins (Figure 2).

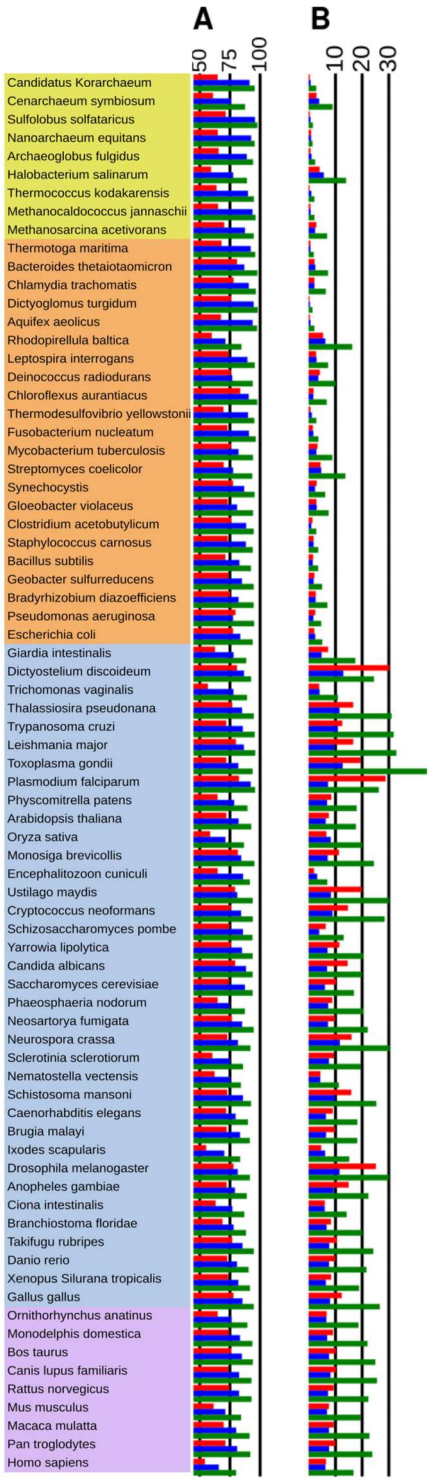


***Figure 2**. Proportion of (A) AR- and (B) EAR-containing proteins per organism predicted by using three predictors of aggregation, ArchCandy 2.0 (red), Pasta 2.0 (blue) and TANGO (green). Archaea, bacteria, eukaryotes and mammalian eukaryotes are outlined by yellow, orange, blue and violet respectively (made by using free options in iTOL, https://itol.embl.de/ (45)).*

Having a global view of the dispersion of aggregation potential of the proteomes, it was interesting to analyze the tendencies associated with groups of the organisms. First, we compared prokaryotes and eukaryotes. All three predictors detect more AR-containing proteins and higher AR-coverage in prokaryotes in comparison to eukaryotes (Figure 3A and 3B). The tendency is reversed when we compare the occurrence of EARs (Figure 3C). The percentage of EAR-containing proteins and coverage of EARs are noticeably higher in eukaryotic than in prokaryotic organisms (Figure 3C and 3D). This can be explained by a higher number of IDRs in eukaryotes, which require the IDRs to mediate a more complex network of protein-protein interactions in comparison to prokaryotes (46, 47). At the same time, the coverage of EARs in IDRs is lower in eukaryotes compared to prokaryotes (Figure 3E). Thus, the eukaryotic IDRs are less aggregation-prone on average than the prokaryotic ones, suggesting a higher selective pressure on their IDRs to avoid aggregation.

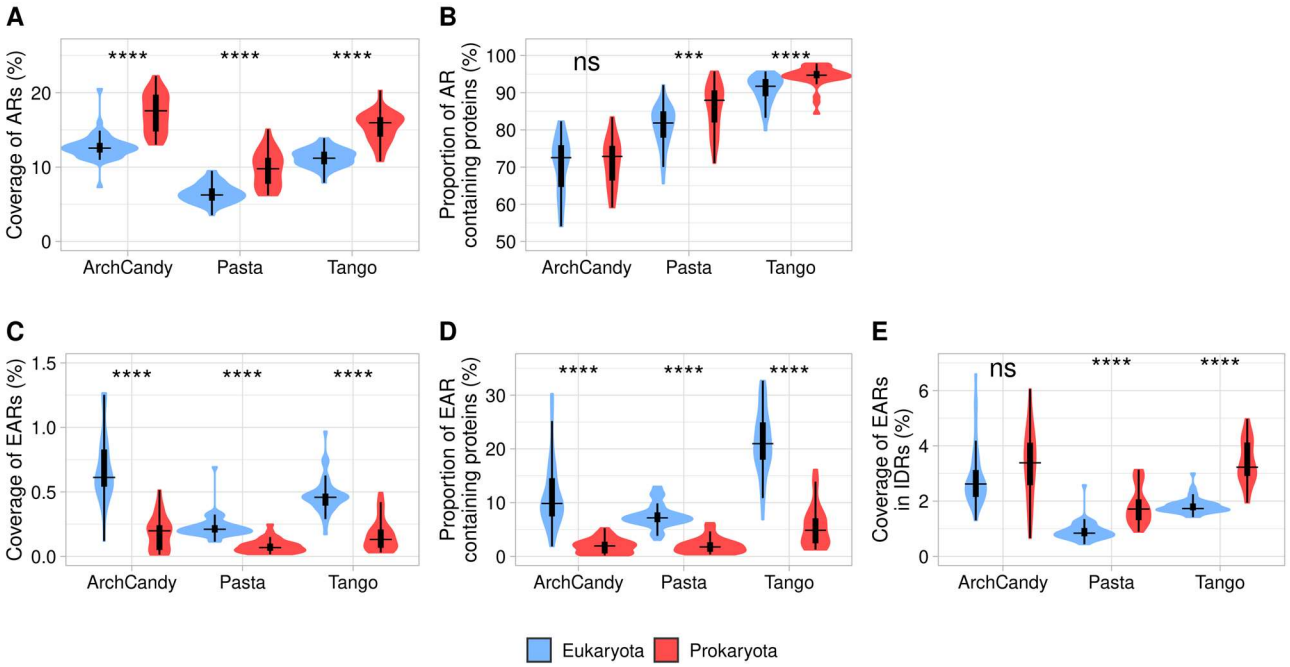


Figure 3. *Level of aggregation potential according to three amyloid predictors in prokaryotes and eukaryotes. Coverage of ARs (A), proportion of AR-containing proteins (B), coverage of EARs (C), proportion of EAR-containing proteins (D) and coverage of EARs in IDRs (E). For statistical analysis between eukaryotic and prokaryotic organisms we performed a t-test for the predictors individually (ns : non-significant; * : p-value < 0.05 ; ** : p-value < 0.01 ; *** : p-value < 0.001 ; **** : p-value < 0.0001).*

The more thermophilic the less aggregation-prone

A unique feature of prokaryotes is the wide range of their optimal growth temperatures (OGTs), some of them reaching temperatures above 105°C (48). We estimated the aggregation potential of the prokaryotic proteomes depending on the OGTs. For this purpose, we subdivided the selected reference proteomes into two groups: 20 mesophilic organisms, those with an OGT below 41°C and 11 thermophilic organisms with an OGT above 41°C. The comparison of proportion and coverage of ARs from these groups do not reach the same conclusion as ArchCandy 2.0 predicts a decrease in ARs in the thermophilic organisms, while PASTA 2.0 and TANGO show the opposite tendency (see Supplementary Figure 1). However, evaluation of EARs by all the predictors clearly demonstrated that they decrease with the increase of OGT (Figure 4). It has been also shown that the frequency of glutamine residue, which has a high amyloidogenic potential, decreases, while the total frequency of charged residues, which can block amyloid-formation, increases in thermophilic proteins(49). At the same time, the temperature increase may favor aggregation. For example,

it has been shown that the amyloidogenesis rate constant of Aβ-peptide increases and the lag time decreases with increasing temperature (50). Considering all this, we can conclude that the decrease in the EARs with OGT can be a result of an evolutionary pressure on the thermophilic proteins to avoid the aggregation.

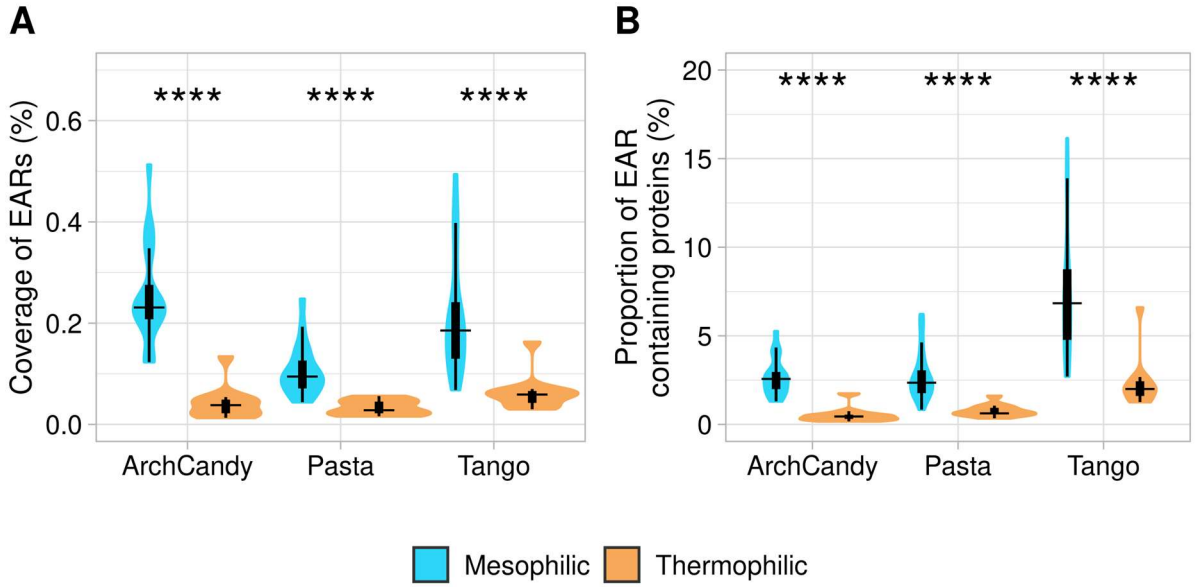


Figure 4. *Coverage of EAR (A) and proportion of EAR-containing proteins (B) in mesophilic (blue) and thermophilic organisms (orange). The analyzed set of thermophilic organisms is: Chloroflexus aurantiacus, Thermodesulfovibrio yellowstonii, Dictyoglomus turgidum, Nanoarchaeum equitans, Sulfolobus solfataricus, Thermotoga maritima, Archaeoglobus fulgidus, Thermococcus kodakaraensis, Methanocaldococcus jannaschii, Candidatus korarchaeum, Aquifex aeolicus. For statistical analysis between mesophilic and thermophilic organisms we performed a t-test for predictors individually (ns : non-significant; * : p-value < 0.05 ; ** : p-value < 0.01 ; *** : p-value < 0.001 ; **** : p-value < 0.0001).*

Occurrence of EARs in proteins depending on their length

In general, the longer the protein chain, the higher the probability for it to have both ARs and EARs. One would expect that if the ARs or EARs are uniformly distributed in protein sequences, their occurrence would correlate linearly with protein length. To see the tendency better, one can normalize the occurrence of ARs/EARs by dividing it by protein length. Previously, similar analyses have been done for the ARs using bacterial proteins (25) and the human proteome (27). Both studies showed that the aggregation potential of a protein normalized by its length goes down with the increase of protein size. To compare this conclusion with our results from the 76 selected proteomes, we analyzed the normalized proportion of AR-containing proteins and normalized AR-coverage depending on length (Figure 5 A, C). In agreement with the previous studies, we observed a decrease in the normalized proportion of AR-containing proteins and AR coverage with length. The steady decrease starts after 500 residues. The graph of AR coverage has a sharp peak at around 350-residue length. Clustering proteins by MMseqs2 (51) at 30 % of sequence identity, we found that this peak contains a significant excess of G protein-coupled receptors having high AR coverage, explaining this anomaly. The 200-500 residue region with the highest AR coverage and proportion coincides with the length ranges where proteins are predicted to be the most structured (Figure 5E) and in general, it negatively correlates with the IDR coverage by protein length. Thus, the AR proportion and coverage curves can be explained by the fact that structured regions have a higher probability of containing ARs, and proteins of less than 500 residues are mostly structured.

The dependence of EARs on protein length demonstrates that it differs from ARs (Figure 5 B, D).

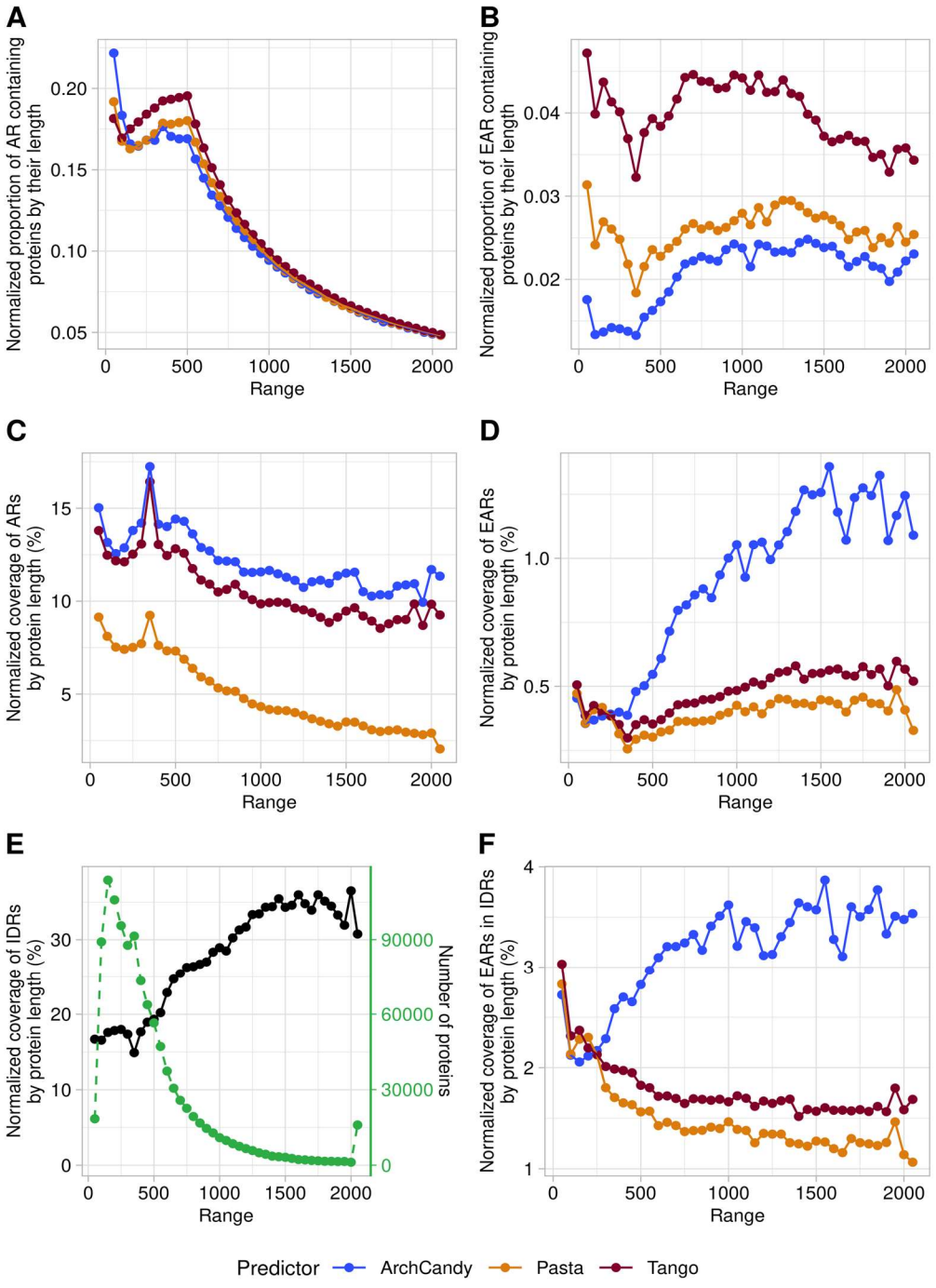


Figure 5. *Plots of the proportion of AR (A) and EAR (B) containing proteins depending on the protein length. Plots of coverage of AR (C) and EAR (D) in proteins according to their length. Plots of coverage of IDR (E) and EAR in IDR (F). Proteins are grouped by subsets of 50 residues (e.g. 1-50, 51-100 etc). Proteins longer than 2000 were grouped into one subset. The predictors used have systematic biases at the terminal regions of proteins and this affects results on the short sequence lengths. To take this bias into account, we also run the predictors against a set of randomized sequences. This set contains proteins from our database with each sequence computationally shuffled, respecting the average amino acid composition of our database and having the same distribution of protein lengths. This allowed us to determine a correction coefficient which was used to adjust the values of EAR, AR and IDR.*

The predictors show a plateau with the lowest EAR-coverage for the shortest proteins (less than 350 residues), which steadily goes up for longer proteins. A similar trend is observed when we plot the dependence of the proportion of EAR-containing proteins by length.
The dependence of the coverage of IDRs against protein length (Figure 5E) is similar to the one of EARs, explaining the low aggregation potential of the short sequences by their tendency to be structured. Indeed, the region of 200-400 residues, which corresponds to the stable structural domains of proteins, has the lowest coverage of IDRs and EARs.
To see the tendency linked only to the characteristics inherent in the IDR sequences, we analyzed the dependence between the EAR coverage in IDRs and the length of proteins. The analysis shows that for TANGO and PASTA 2.0, shorter sequences have higher EAR coverage in IDRs. In contrast, ArchCandy 2.0 predicted an increase of EAR coverage in IDRs with protein length (Figure 5F). One explanation of this discrepancy between the predictors maybe the fact that ArchCandy predicts Asn/Gln-rich regions, which are frequently found in long proteins, as aggregation prone, while TANGO and PASTA do not.
Thus, we do not observe a decrease in aggregation potential with an increase of protein size when we consider EARs. The longer a protein chain, the higher its propensity to aggregate. Therefore, the question arises as to the mechanism preventing fibril formation of long proteins. One possible explanation can be that long proteins, having multiple IDRs, represent “steric brushes”, preventing their intermolecular interactions and aggregation due to the high entropic barrier (52).

Occurrence of EAR-containing proteins in different cellular compartments

Proteins having different cellular localizations may differ in their aggregation potential. Therefore, we analyzed the occurrence of AR- and EAR-containing proteins in 4 major subcellular localizations: secreted proteins identified by SignalP (40), transmembrane proteins by using TMHMM (39), nuclear proteins with NLS (nuclear localization signals) found by SLiMs (41), and the remaining proteins that were considered mostly cytosolic (Figure 6). We observed similar levels of AR-containing proteins in all compartments except the transmembrane proteins, which have significantly higher levels (Supplementary Figure 2). The high level of AR-containing proteins among the transmembrane proteins was expected because their hydrophobic TM helices are detected as ARs by all predictors. The most striking observation was the high level of EAR-containing proteins in nucleus of eukaryotes, which is at least two times higher than in the other cellular localizations (Figure 6A). In line with this result, it has been shown previously that under stress conditions, proteins in the nucleus tend to form aggregates (53).
In prokaryotes, we observed more EAR-containing proteins among those involved in the secretory pathway in comparison to those present in the transmembrane and cytosol (Figure 6B). This tendency suggests that the secreted proteins being outside of the cell are under a reduced evolutionary pressure to avoid aggregation. Formation of aggregates out of the cell may be less deleterious for unicellular prokaryotic organisms in comparison with most of the eukaryotes, which can accumulate unwanted deposits within the extracellular space of their tissues. Moreover, it is known that many prokaryotes use secreted proteins to form functional amyloids (5).

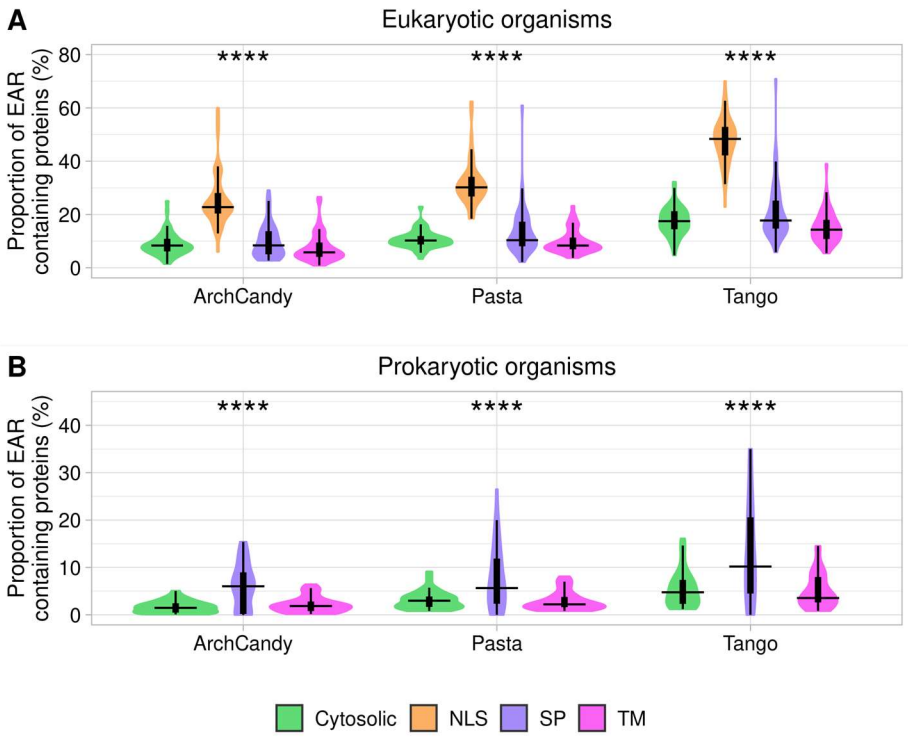


Figure 6. *Plots of proportion of EAR-containing proteins according to the protein localization in eukaryotic (A) and prokaryotic (B) organisms. Proteins are splitted in four groups: cytosolic proteins, extracellular ones with signal peptides (SP), transmembrane proteins (TM) and nuclear proteins having nuclear localization signals (NLS). The proportions of the analyzed proteins are: in eukaryotes, 58.5% are cytosolic, 6.2% have SPs, 21.4% are transmembrane proteins and 13.9% have NLS. In prokaryotes, 71.9 % are cytosolic, 4.3% have SPs and 23.7% are transmembrane proteins. For statistical analysis between the different cell compartments we performed an anova test for the predictors individually (ns : non-significant; * : p-value < 0.05 ; ** : p-value < 0.01 ; *** : p-value < 0.001 ; **** : p-value < 0.0001).*

Relationship between cellular abundance of proteins and ARs/EARs frequencies

The amount of genome-wide data on gene expression has drastically increased in the past few years (54). The data comes from various technologies, organisms, and tissues (normal or disease related), making it difficult to compare them in a large-scale analysis. In this case, we find that the data from the Protein Abundance Database (PaxDb) (55) are the most suitable for our purposes. PaxDb represents protein abundance by “protein per million” (ppm) and by doing so, overcomes the problem of variability in cell size or dilutions in the samples used, making comparisons between them possible. The PaxDb has the protein expression level in different tissues and organs of organisms. Additionally, it provides the average abundance of a protein in the whole organism. We used this average abundance value to analyze the expression level of AR-/EAR-containing proteins, which are available both in PaxDb and in our dataset.

Expression levels range from almost zero up to more than 100 000 ppm. The majority of proteins have values of less than 2 ppm. The number of proteins with the abundance more than 50 ppm drops significantly (Figure 7). Therefore, we grouped these proteins together in our analysis.

The analysis revealed that the frequency of occurrence of EAR-containing proteins decreases with the ppm growth and is becoming lower than non-EAR-containing proteins. From the observed dependence of the difference between EAR- and non-EAR-containing proteins, we can conclude that highly expressed

proteins are less prone to aggregate, with this finding being consistent in the three predictors used. We observed a similar tendency with the frequency of occurrence of AR-containing proteins depending on the abundance (see Supplementary Figure 3). It suggests that highly expressed proteins are under a greater selective pressure to avoid aggregation. This conclusion is in agreement with a previous study of human proteins also suggesting that aggregation-prone proteins and gene level expression are inversely correlated (56).

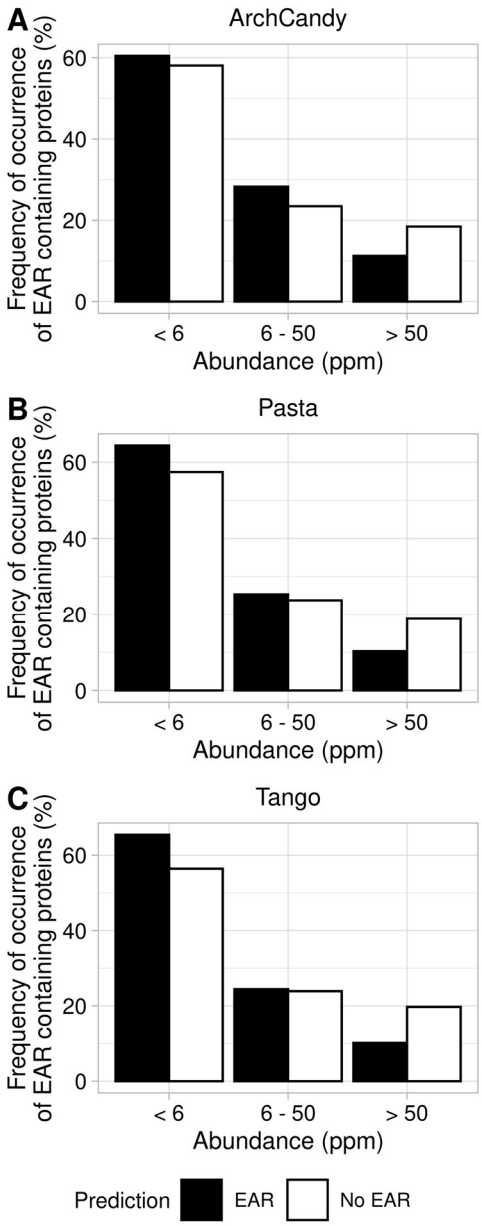


***Figure 7**. Frequency of occurrence of EAR -containing proteins predicted by ArchCandy 2.0 (A), Pasta 2.0 (B) and TANGO (C). Proteins are grouped based on their abundance in three groups : less than 5 ppm, 5-50 ppm, and more than 50 ppm.*

EAR levels in essential proteins

As demonstrated previously, essential genes are subject to a greater selection pressure than non-essential genes (55, 57). It has also been shown that essential proteins are less prone to aggregation (20, 23). In order to find essential proteins in our database, we used the DEG database of essential proteins (58) and run BLAST program with E-value < 0,001(59). By this approach we identified 705692 essential proteins (~62,6 %) in our database. Analysis of these proteins by the three predictors showed a lower EAR coverage, and in a lesser extent, proportion of essential and non-essential EAR-containing proteins in eukaryotes (Figure 8 A, C). Our results are in agreement with previous conclusions that essential proteins have a lower aggregation score than non-essential proteins (23). In prokaryotes, we observe the opposite tendency (Figure 8B, D). Previous analyses of ARs (not EARs) made on a smaller scale in bacteria (25) have shown that essential proteins have less ARs. Our results of the AR analysis in prokaryotes (see Supplementary Figure 4) are in agreement with this conclusion.

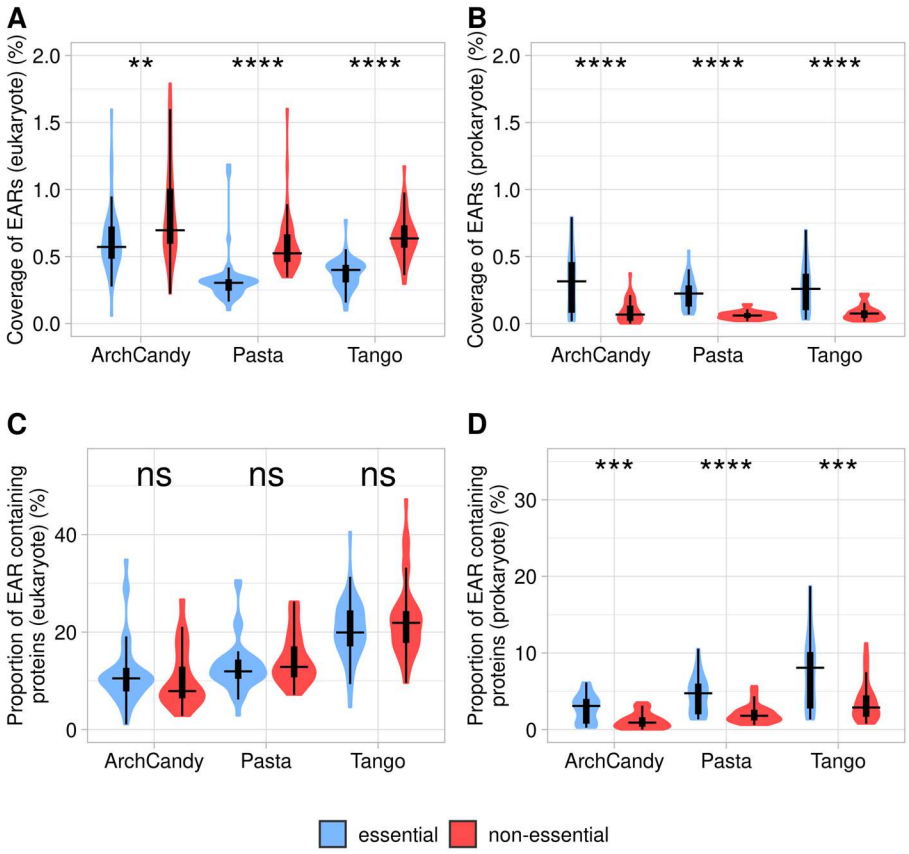


Figure 8. *Coverage of EARs in essential and non-essential proteins in eukaryote (A) and prokaryote (B) organisms. Proportion of EAR-containing proteins known as essential or non-essential in eukaryote (C) and prokaryote (D) organisms. For statistical analysis between essential and non-essential proteins we performed a t-test for amyloidogenic predictors individually (ns : non-significant; * : p-value < 0.05 ; ** : p-value < 0.01 ; *** : p-value < 0.001 ; **** : p-value < 0.0001).*

Short Linear Motifs (SLiMs) in EARs

A significant portion of protein interactions are mediated by short linear motifs (SLiMs) preferentially found in IDRs (41). As EARs are also located within the IDRs, it was interesting to analyze the co-occurrence of SLiMs and EARs in proteins. Although both prokaryotes and eukaryotes have functional SLiMs, the

eukaryotic linear motifs are more common, as well as better classified and documented. Most of the eukaryotic SLiMs can be found in the ELM resource (41) alongside with their descriptions, experimental evidence from the literature, and Regular Expressions (RegEx) of the recurrent patterns. Therefore, we focused our analysis on the SLiMs from eukaryotes. For this purpose, we applied the RegEx from the ELM database (41) to the IDRs and EARs determined by our pipeline (33). The SLiMs are subdivided into 6 major classes: (LIG) ligand binding motifs and (DOC) docking sites both involved in protein-protein interactions of the functional complexes, (MOD) modification sites covering several post-translational modifications of proteins (e.g. phosphorylation, palmitoylation, glucosylation), (DEG) sites of proteins that are important in regulation of protein degradation rates, (TRG) targeting sites responsible for protein sorting in cellular compartments and (CLV) specific cleavage sites.

The results of all three aggregation predictors showed that EAR-containing proteins are enriched in SLiMs in comparison to IDR-containing proteins without EARs (Figure 9). By using the exact Fisher test, we were able to select SLiMs, which are significantly enriched in EAR- containing proteins compared to IDR-containing proteins alone. Interestingly, 20 of the 25 degradation motifs (proteasome pathway) from DEG class occur more frequently in EAR-containing proteins than in non-EAR-containing proteins (Figure 9). 17 of the 22 TRGs are also more frequently present in EAR-containing proteins than in IDR-containing proteins. Among them, 3 SLiMs were found to be Endosome-Lysosome-Basolateral sorting signals. These results suggest that EAR-containing proteins may be more susceptible to being degraded by proteasome and lysosome pathways compared to just IDR-containing proteins. This might be a strategy used by organisms to prevent protein aggregation by increasing the degradation of potential aggregation-prone proteins. Cleavage sites (CLV) are less prevalent in EARs, which may prevent the release of smaller amyloidogenic peptides such as the well-known Aβ-peptide (60).

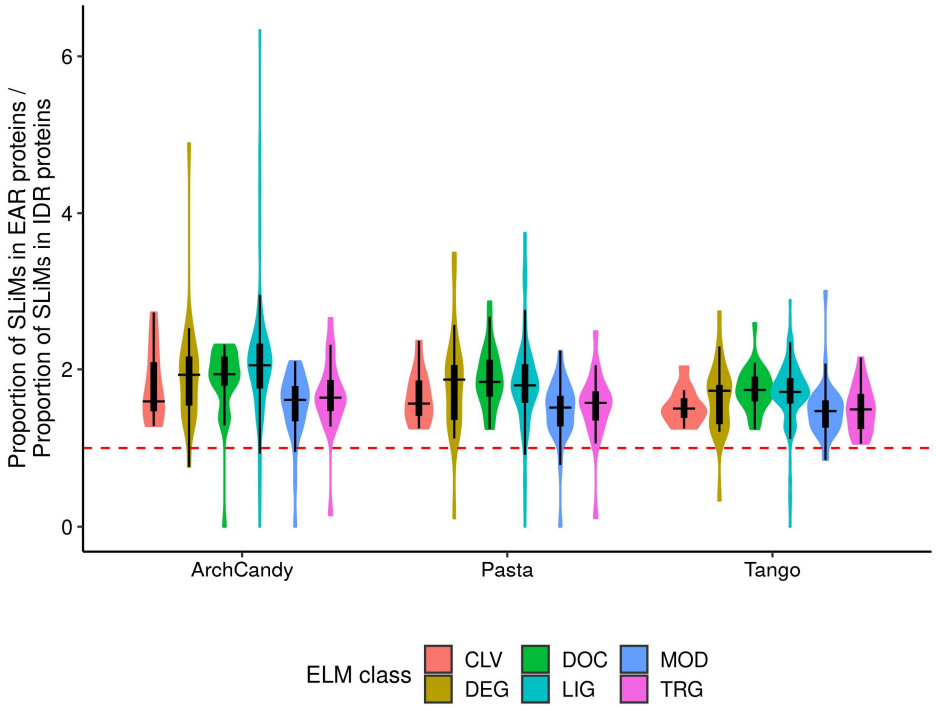


Figure 9. *Ratio of proportions of SLiMs in EAR-containing proteins and in IDR-containing proteins without EAR, predicted by three predictors. Each dot represents a given SLiMs grouped in 6 classes denoted by different colors. The majority of the SLiMs have their ratios greater than 1.0 (red dotted line), meaning that they are enriched in EAR-containing proteins.*

Functional domains enriched in EAR-containing proteins

With a method similar to the SliMs enrichment, we tried to identify functional Pfam domains enriched in EAR-containing proteins. Of the 15116 known Pfam domains, only 1410 are significantly more prevalent in EAR-containing proteins predicted by ArchCandy 2.0 (exact Fisher test, p-value < 0.001). 484 of them belong to 154 clans according to the classification of Pfam. The functional domains and clans that came on top are: nucleoporin FG repeat region (CL0647), RNA recognition motif domains (CL0221) and zinc-finger domains (CL0511, CL0390) (see Supplementary data 1). We searched the experimental evidence of aggregation by these domains in the literature and found that the nucleoporin proteins are known to form amyloids(61). EAR-containing proteins predicted by both ArchCandy 2.0 and TANGO are positively enriched in nucleoporin FG repeat region (CL0647). From the known functional amyloids described in the literature, we also found back RIPK1 and RIPK3 (8, 9) and PMEL17 (62), which were conserved in 6 distinct proteins from mammalians with the prediction of ArchCandy 2.0 but not Pasta 2.0 or TANGO.

Previous studies of Pfam domains and gene ontology (GO) term enrichment in amyloidogenic proteins (24, 28) pointed out the over-representation of membrane transport activity, pH and ion regulation and even cytoskeleton organization. However, they considered ARs not EARs. Therefore, we did not find most of the afore mentioned functions in our analysis.

Conservation of EARs sequences

Another approach to find new functional amyloids is to search for EARs that are conserved among different species. For this purpose, we reduced EARs predicted by either ArchCandy 2.0, TANGO or Pasta 2.0 with CD-HIT (63) at 70 % sequence identity and 90 % of coverage, to obtain a non-redundant set of the EARs (Table 1). Then, we ran BLAST (59) for each EAR sequence against all proteins from our redundant database to select only conserved EAR sequences (Figure 10). This gave us for each EAR a Multiple Sequence Alignment (MSA) of similar sequences found in other proteins. Some sequences of the MSA were EARs and the others were not according to the predictors. We selected the MSAs with EARs in more than five other proteins and further reduced the MSA number by merging those that shared at least 80 % of the same conserved EAR. This clustering results found 2218, 869 and 178 of the most conserved EAR sequences for ArchCandy 2.0, Pasta 2.0 and TANGO, respectively (Table 1). We observed that only a small number of EAR sequences are conserved out of more than one million proteins. Among them we found already known functional amyloids, such as RIPK3 and RIPK1 (8) and PMEL17 (62). This suggests that the list of conserved EARs found by this protocol (Supplementary data 2) can be used for detection and experimental tests of new functional amyloids.

Table 1. Number of EARs at each step of the protocol.

Predictor	Number of non-redundant EARs	EARs found one time in MSA	EARs found 2 to 5 times in MSA	EARs found more than 5 times in MSA	Number of clusters with the most conserved EARs
ArchCandy 2.0	93229	72153 (77.4%)	16124 (17.3%)	4952 (5.3%)	2218
Pasta 2.0	42997	35412 (82.4%)	5683 (13.2%)	1902 (4.4%)	869
TANGO	13816	12342 (89.3%)	1219 (8.8%)	255 (1.8%)	178

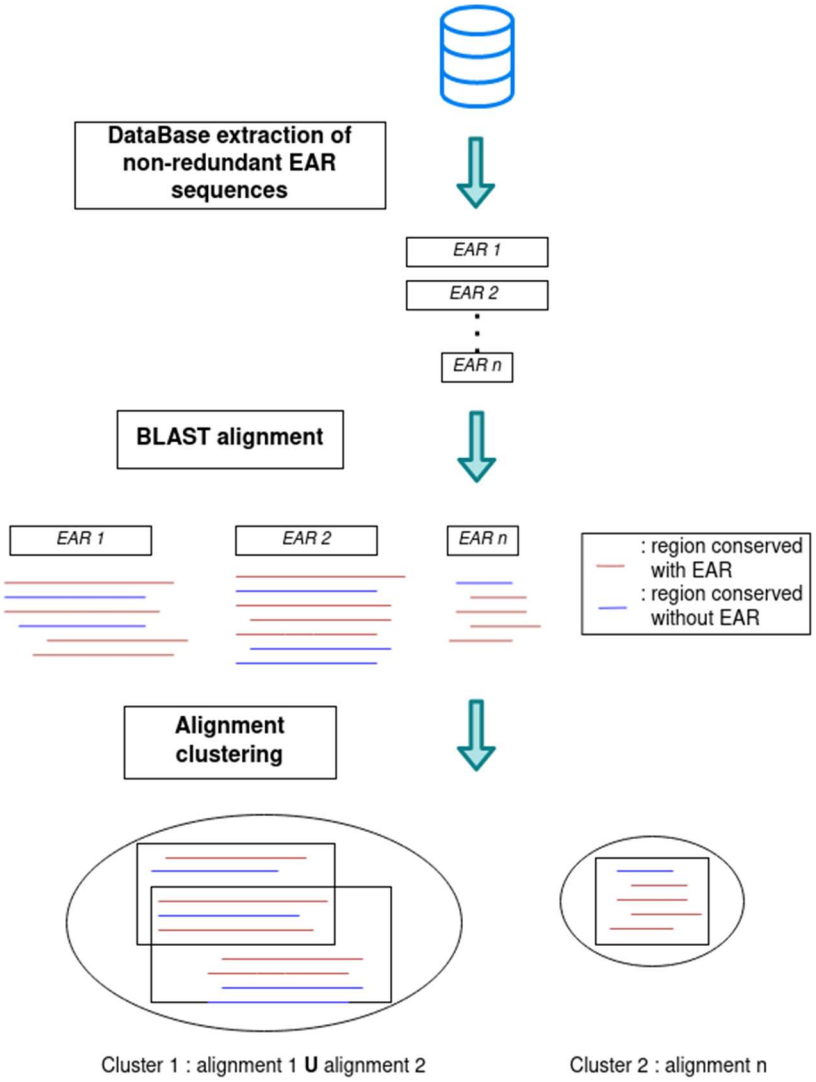


***Figure 10**. Protocol for the evaluation of EAR sequence conservation.*

DISCUSSION

The recent progress with computational approaches predicting aggregation (10, 12, 13, 15, 16, 18, 31, 33), and an increasing number of whole-proteome sequencing data, opened an avenue for the comprehensive census of aggregation-prone regions in proteins. In this work, we performed the detailed analysis of 76 full reference proteomes from the UniProt databank. As a result, a number of interesting correlations, confirmed by all the predictors used in this work (ArchCandy 2.0, Pasta 2.0 and TANGO), were discovered. First, we detected a significantly lower percentage of EAR-containing proteins (about 10%) in comparison with a high percentage of AR-containing proteins in proteomes (about 80%). The number of EARs correlates better with a small number of the known proteins forming aggregates *in vivo*, and, therefore, EARs can be suggested as a more precise measure of the aggregation potential of proteins. We showed that there are more ARs in prokaryotes than in eukaryotes and that this tendency is inverted for EARs. Second, we found that the thermophilic prokaryotes have significantly less EARs and ARs in comparison to mesophilic prokaryotes. The correlation may reflect an evolutionary pressure on the thermophilic proteins, because the amyloid formation rate constant increases with temperature (50).

Additionally, in agreement with previous studies, we observed small decrease in the normalized AR

coverage with protein length. However, we do not observe a decrease in the aggregation potential of sequences with an increase of protein size when we consider EARs. In our opinion, the mechanism of prevention of aggregation of long proteins has an entropic basis, where the other parts of the chain generate repulsive forces for intermolecular interactions similar to molecular brushes.

It worth mentioning that our analysis did not confirm previously published conclusions that the average aggregation propensity of a proteome correlates inversely with the complexity and longevity of the studied organisms (29).

It was also shown that proteins having different cellular localizations differ in the aggregation potential. For example, the level of EAR-containing proteins in nuclear proteins of eukaryotes is about two times higher than in the other cellular localizations. In prokaryotes, we observed more EAR-containing proteins among those involved in the secretory pathway in comparison to the transmembrane and cytosolic proteins. This tendency suggests that the secreted proteins being outside of the cell are under a reduced evolutionary pressure to avoid aggregation. Remarkably, a great majority of EAR-containing proteins are enriched in SLiMs in comparison to IDR-containing proteins without EARs. We also noticed that highly expressed proteins are less prone to aggregate suggesting that highly expressed proteins are under a greater negative selective pressure in order to avoid the aggregation. Finally, we revealed a greater level of aggregation predicted in non-essential proteins compared to essential proteins.

Thus, we performed the census of the aggregation-prone regions in proteomes. A number of new relationships found in this work led us to a better understanding of the functional and evolutionary relations of protein aggregation in organisms from the three kingdoms of life: eukaryote, bacteria and archaea. Beyond this, our study opens up new opportunities for a number of experimental tests.

ACKNOWLEDGMENT

The authors thank Priya Amin for her assistance with the English.

FUNDING

This work was supported by REFRACT project with Latin America in RISE program (2018-2023) H2020-MSCA-RISE-2018 to A.V.K.; Azerbaijan National Academy of Sciences and The Ministry of Science and Education of Azerbaijan to Z.O.; the Ministère de l'Education Nationale de la Recherche et de Technologie (MENRT) to E.V. and F.R.; by a CNRS PhD fellowship to T.F.

Author Contributions: Conceptualization, A.V.K., T.F. and E.V.; methodology, A.V.K., T.F. and E.V.; software, T.F. and E.V.; data curation, T.F., F.R. and Z.O.; writing—original draft preparation, T.F. and A.V.K.; writing—review and editing, A.V.K., T.F., Z.O, F.R, E.V.; supervision, A.V.K.

CONFLICT OF INTEREST: Authors declare no conflict of interest.

REFERENCES

1. Steven, A.C., Baumeister, W., Johnson, L.N., & Perham,R.N. (2016) Molecular Biology of Assemblies and Machines. *Garl. Sci.*, **1**, 5–24.

2. Benson,M.D., Buxbaum,J.N., Eisenberg,D.S., Merlini,G., Saraiva,M.J.M., Sekijima,Y., Sipe,J.D. and Westermark,P. (2020) Amyloid nomenclature 2020: update and recommendations by the International Society of Amyloidosis (ISA) nomenclature committee. *Amyloid*, **27**, 217–222.

3. Prusiner,S.B. (1998) Prions. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 13363–13383.

4. Bondarev,S.A., Antonets,K.S., Kajava,A. V, Nizhnikov,A.A. and Zhouravleva,G.A. (2018) Protein co-aggregation related to amyloids: Methods of investigation, diversity, and classification. *Int. J. Mol. Sci.*, **19**, 1–30.

5. Erskine,E., MacPhee,C.E. and Stanley-Wall,N.R. (2018) Functional Amyloid and Other Protein Fibers in the Biofilm Matrix. *J. Mol. Biol.*, **430**, 3642–3656.

6. Greenwald,J. and Riek,R. (2010) Biology of amyloid: Structure, function, and regulation. *Structure*, **18**, 1244–1260.

7. Barnhart,M.M. and Chapman,M.R. (2006) Curli biogenesis and function. *Annu. Rev. Microbiol.*, **60**, 131–147.

8. Kajava,A. V, Klopffleisch,K., Chen,S. and Hofmann,K. (2014) Evolutionary link between metazoan RHIM motif and prion-forming domain of fungal heterokaryon incompatibility factor HET-s/HET-s. *Sci. Rep.*, **4**, 1–6.

9. Li,J., McQuade,T., Siemer,A.B., Napetschnig,J., Moriwaki,K., Hsiao,Y.S., Damko,E., Moquin,D., Walz,T., McDermott,A., *et al.* (2012) The RIP1/RIP3 necrosome forms a functional amyloid signaling complex required for programmed necrosis. *Cell*, **150**, 339–350.

10. Ahmed,A.B., Znassi,N., Château,M.T. and Kajava,A. V (2015) A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's Dement.*, 10.1016/j.jalz.2014.06.007.

11. Ahmed,A.B. and Kajava,A. V (2013) Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence. *FEBS Lett.*, **587**, 1089–1095.

12. Conchillo-Solé,O., de Groot,N.S., Avilés,F.X., Vendrell,J., Daura,X. and Ventura,S. (2007) AGGRESCAN: A server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics*, **8**.

13. Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.

14. Tartaglia,G.G., Pawar,A.P., Campioni,S., Dobson,C.M., Chiti,F. and Vendruscolo,M. (2008) Prediction of Aggregation-Prone Regions in Structured Proteins. *J. Mol. Biol.*, **380**, 425–436.

15. Thompson,M.J., Sievers,S.A., Karanicolas,J., Ivanova,M.I., Baker,D. and Eisenberg,D. (2006) The 3D profile method for identifying fibril-forming segments of proteins. *PNAS March*, **14**, 4074–4078.

16. Walsh,I., Seno,F., Tosatto,S.C.E. and Trovato,A. (2014) PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.*, 10.1093/nar/gku399.

17. Louros,N., Orlando,G., De Vleeschouwer,M., Rousseau,F. and Schymkowitz,J. (2020) Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. Commun.*, **11**, 1–13.

18. Wojciechowski,J.W. and Kotulska,M. (2020) PATH - Prediction of Amyloidogenicity by Threading and Machine Learning. *Sci. Rep.*, **10**, 1–9.

19. Antonets,K.S., Kliver,S.F. and Nizhnikov,A.A. (2018) Exploring Proteins Containing Amyloidogenic Regions in the Proteomes of Bacteria of the Order Rhizobiales. *Evol. Bioinforma.*, **14**.

20. Tartaglia,G.G. and Vendruscolo,M. (2009) Correlation between mRNA expression levels and protein

aggregation propensities in subcellular localisations. *Mol. Biosyst.*, **5**, 1873–1876.

21. Antonets,K.S. and Nizhnikov,A.A. (2017) Predicting amyloidogenic proteins in the proteomes of plants. *Int. J. Mol. Sci.*, **18**.

22. Castillo,V., Graña-Montes,R., Sabate,R. and Ventura,S. (2011) Prediction of the aggregation propensity of proteins from the primary sequence: Aggregation properties of proteomes. *Biotechnol. J.*, **6**, 674–685.

23. Chen,Y. and Dokholyan,N. V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.*, **25**, 1530–1533.

24. Das,S., Pal,U., Das,S., Bagga,K., Roy,A., Mrigwani,A. and Maiti,N.C. (2014) Sequence complexity of amyloidogenic regions in intrinsically disordered human proteins. *PLoS One*, **9**.

25. De Groot,N.S. and Ventura,S. (2010) Protein aggregation profile of the bacterial cytosol. *PLoS One*, **5**.

26. Goldschmidt,L., Teng,P.K., Riek,R. and Eisenberg,D. (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 3487–3492.

27. Monsellier,E., Ramazzotti,M., Taddei,N. and Chiti,F. (2008) Aggregation propensity of the human proteome. *PLoS Comput. Biol.*, **4**.

28. Prabakaran,R., Goel,D., Kumar,S. and Gromiha,M.M. (2017) Aggregation prone regions in human proteome: Insights from large-scale data analyses. *Proteins Struct. Funct. Bioinforma.*, **85**, 1099–1118.

29. Tartaglia,G.G., Pellarin,R., Cavalli,A. and Caflisch,A. (2005) Organism complexity anti-correlates with proteomic $β$-aggregation propensity. *Protein Sci.*, **14**, 2735–2740.

30. Pawar,A.P., DuBay,K.F., Zurdo,J., Chiti,F., Vendruscolo,M. and Dobson,C.M. (2005) Prediction of ‘aggregation-prone’ and ‘aggregation- susceptible’ regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.*, **350**, 379–392.

31. Tartaglia,G.G. and Vendruscolo,M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.*, **37**, 1395–1401.

32. Villain, Etienne,A.A.N. and Kajava,A. V (2018) Porins and amyloids are coded by similar sequence motifs Etienne. *Proteomics*, 10.1002/jssc.201200569.

33. Falgarone,T., Villain,É., Guettaf,A., Leclercq,J. and Kajava,A. V. (2022) TAPASS: Tool for annotation of protein amyloidogenicity in the context of other structural states. *J. Struct. Biol.*, **214**.

34. Cao,Q., Boyer,D.R., Sawaya,M.R., Ge,P. and Eisenberg,D.S. (2019) Cryo-EM structures of four polymorphic TDP-43 amyloid cores. *Nat. Struct. Mol. Biol.*, **26**, 619–627.

35. Bateman,A. (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506--D515.

36. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

37. Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2017)

CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.

38. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *Cit. Eddy SR*, **7**, 1002195.

39. Krogh,A., Larsson,B., Von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

40. Petersen,T.N., Brunak,S., Von Heijne,G. and Nielsen,H. (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

41. Kumar,M., Gouw,M., Michael,S., Sámano-Sánchez,H., Pancsa,R., Glavina,J., Diakogianni,A., Valverde,J.A., Bukirova,D., Signalyševa,J., *et al.* (2020) ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.*, **48**, D296--D306.

42. Ruhanen,H., Hurley,D., Ghosh,A., O'Brien,K.T., Johnston,C.R. and Shields,D.C. (2014) Potential of known and short prokaryotic protein motifs as a basis for novel peptide-based antibacterial therapeutics: A computational survey. *Front. Microbiol.*, **5**, 1–18.

43. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427--D432.

44. Mier,P., Paladin,L., Tamana,S., Petrosian,S., Hajdu-Soltész,B., Urbanek,A., Gruca,A., Plewczynski,D., Grynberg,M., Bernadó,P., *et al.* (2020) Disentangling the complexity of low complexity proteins. *Brief. Bioinform.*, **21**, 458–472.

45. Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

46. Pancsa,R. and Tompa,P. (2012) Structural disorder in eukaryotes. *PLoS One*, **7**.

47. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.*, **337**, 635–645.

48. Stetter,K.O. (2006) History of discovery of the first hyperthermophiles. *Extremophiles*, **10**, 357–362.

49. Villain,E., Fort,P. and Kajava,A. V (2022) Aspartate-phobia of thermophiles as a reaction to deleterious chemical transformations. *BioEssays*, **44**, 2100213.

50. Tiiman,A., Krishtal,J., Palumaa,P. and Tõugu,V. (2015) In vitro fibrillization of Alzheimer's amyloid-β peptide (1-42). *AIP Adv.*, **5**.

51. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

52. Rubinstein, M and Colby,R. (2003) Polymer Physics, Oxford University Press. *New York.*

53. Karamanos,T.K., Kalverda,A.P., Thompson,G.S. and Radford,S.E. (2015) Mechanisms of amyloid formation revealed by solution NMR. *Prog. Nucl. Magn. Reson. Spectrosc.*, **88**–**89**, 86–104.

54. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big data: Astronomical or genomical? *PLoS Biol.*, **13**, 1–11.

55. Wang,M., Herrmann,C.J., Simonovic,M., Szklarczyk,D. and von Mering,C. (2015) Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**, 3163–3168.

56. Tartaglia,G.G. and Vendruscolo,M. (2009) Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. Biosyst.*, **5**, 1873–1876.

57. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E. V (2002) Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Res.*, **12**, 962–968.

58. Luo,H., Lin,Y., Liu,T., Lai,F.L., Zhang,C.T., Gao,F. and Zhang,R. (2021) DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Res.*, **49**, D677–D686.

59. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

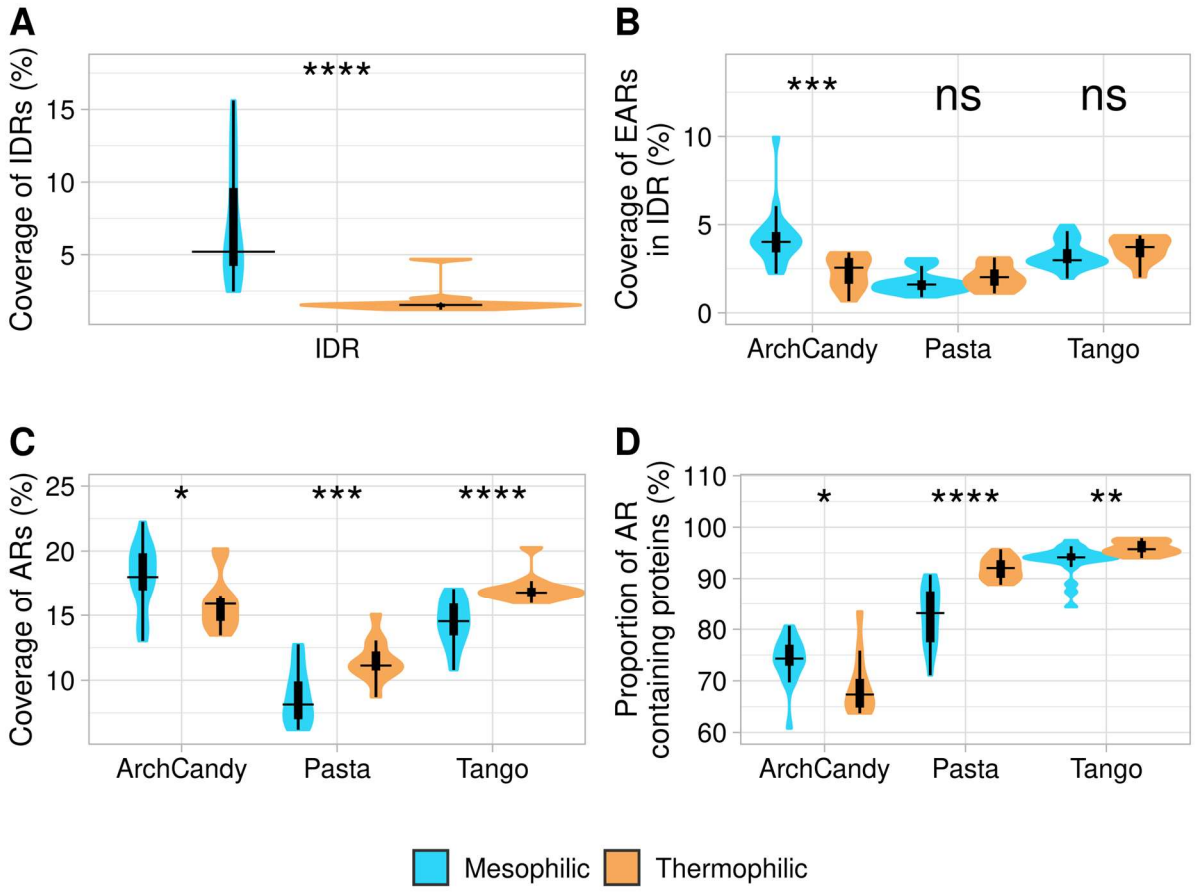
60. Lu,D.C., Rabizadeh,S., Chandra,S., Shayya,R.F., Ellerby,L.M., Ye,X., Salvesen,G.S., Koo,E.H. and Bredesen,D.E. (2000) A second cytotoxic proteolytic peptide derived from amyloid β-protein precursor. *Nat. Med.*, **6**, 397–404.

61. Danilov,L.G., Moskalenko,S.E., Matveenko,A.G., Sukhanova,X. V, Belousov,M. V, Zhouravleva,G.A. and Bondarev,S.A. (2021) The human nup58 nucleoporin can form amyloids in vitro and in vivo. *Biomedicines*, **9**, 1–12.

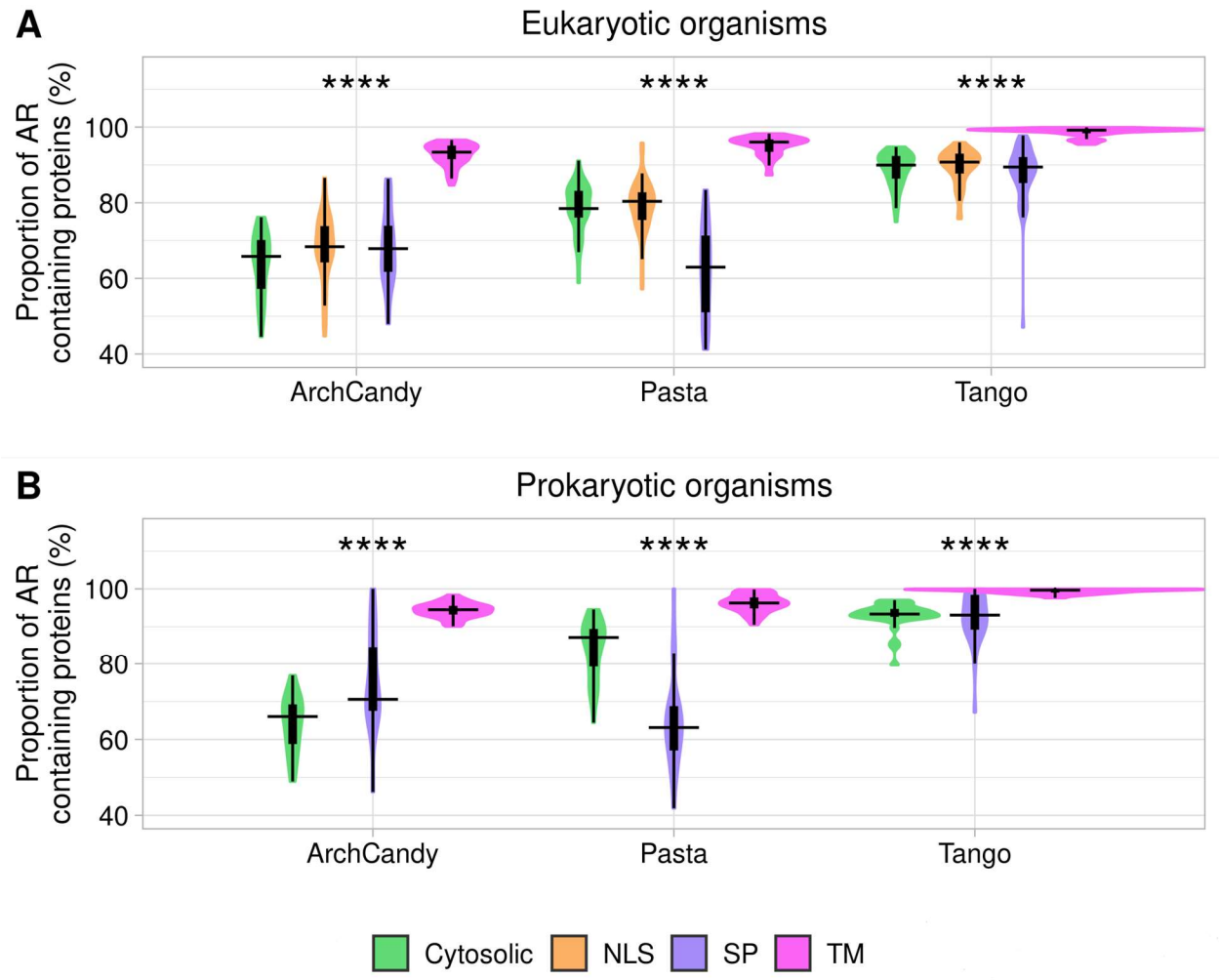
62. Raposo,G. and Marks,M.S. (2002) The dark side of lysosome-related organelles: Specialization of the endocytic pathway for melanosome biogenesis. *Traffic*, **3**, 237–248.

63. Li,W. and Godzik,A. (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

SUPPLEMENTARY DATA

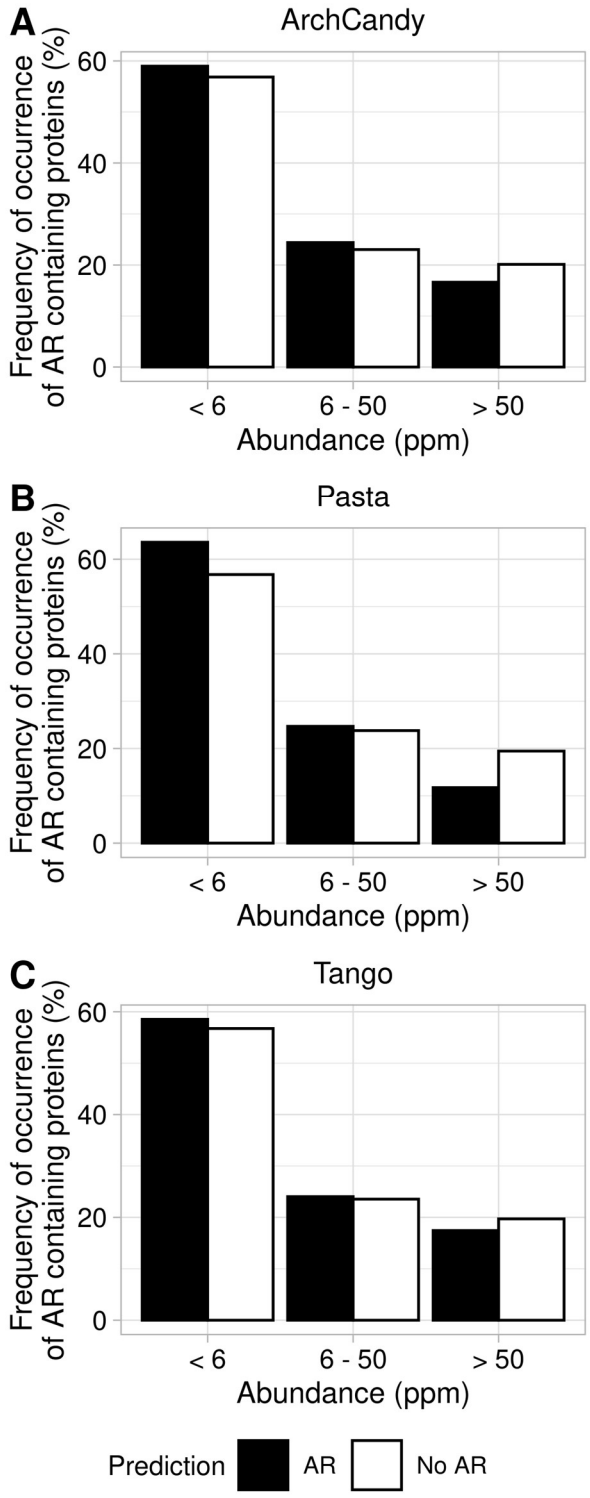


Supplementary Figure 1. Coverage of IDR (A), EAR in IDR (B), AR (C) and proportion of AR containing proteins (D) in mesophilic (blue) and thermophilic organisms (red). The analyzed set of thermophilic organisms is : Chloroflexus aurantiacus, Thermodesulfovibrio yellowstonii, Dictyoglomus turgidum, Nanoarchaeum equitans, Sulfolobus solfataricus, Thermotoga maritima, Archaeoglobus fulgidus, Thermococcus kodakaraensis, Methanocaldococcus jannaschii, Candidatus korarchaeum, Aquifex aeolicus.

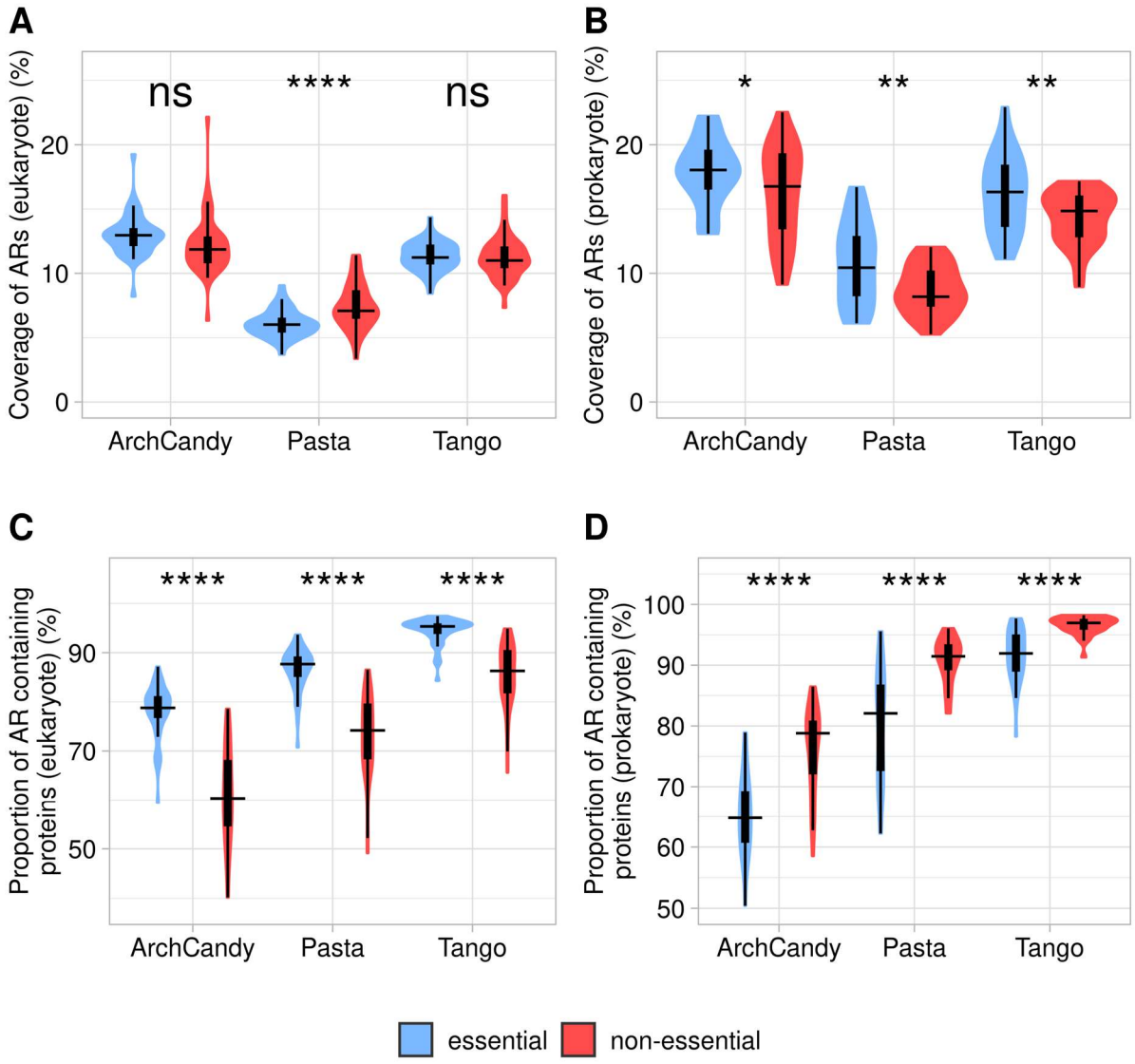


Supplementary Figure 2.

Plots of proportion of AR containing proteins according to the protein localization in eukaryotic (A) and prokaryotic (B) organisms. Proteins are splitted in four groups: cytosolic proteins, extracellular ones with signal peptides (SP), transmembrane proteins (TM) and nuclear proteins having nuclear localization signals (NLS). The proportions of the analyzed proteins are: in eukaryotes, 58.5% are cytosolic, 6.2% have SPs, 21.4% are transmembrane proteins and 13.9% have NLS. In prokaryotes, 71.9 % are cytosolic, 4.3% have SPs and 23.7% are transmembrane proteins



Supplementary Figure 3. Frequency of occurrence of AR containing proteins predicted by ArchCandy 2.0 (A), Pasta 2.0 (B) and TANGO (C). Proteins are grouped based on their abundance in ranges of 5 ppm, proteins with 50 ppm or more are grouped in one range.



Supplementary Figure 4. Coverage of AR in essential and non-essential proteins in eukaryote (A) and prokaryote (B) organisms. Proportion of AR containing proteins known as essential or non-essential in eukaryote (C) and prokaryote (D) organisms.