Why do eukaryotic proteins contain more intrinsically disordered regions?

Walter Basile[1,2], Arne Elofsson[1,2,3,*]

1 Science for Life Laboratory, Stockholm University SE-171 21 Solna, Sweden
2 Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
3 Swedish e-Science Research Center (SeRC)

*** Corresponding author: arne@bioinfo.se**

Abstract

Intrinsic disorder is an important aspect in eukaryotic proteins. However, it is not clear why intrinsic disorder is significantly more frequent in eukaryotic proteins than in prokaryotic proteins. Here, we show that the difference in intrinsic disorder can largely be explained by an increase of serines, primarily in linker regions. Eukaryotic proteins contain about 8% serine, while prokaryotic proteins have roughly 6%. Serine is one of the most disorder-promoting residues and is particularly frequent in the linker regions, connecting domains in eukaryotic proteins. In addition to being disorder-promoting serine is one of the targets for serine/threonine kinases, a large family of predominantly eukaryotic proteins. Phosphorylation often results in a functional change of the target, affecting a multitude of cellular processes, including division, proliferation, apoptosis, and differentiation. However, tyrosine and threonine, the other substrates for this family of kinases, are not more frequent in eukaryotes than prokaryotes. The consistently increased serine frequency suggests that there exist a selective process that promotes intrinisic disorder in eukaryotic proteins. It is possible that one driving force is the increased need for regulation of protein activities through the phosphorylation of serine residues.

Introduction

To deal with the increased complexity in a eukaryotic cell, the eukaryotic proteomes have evolved to differ significantly from prokaryotic ones. The most notable difference is that eukaryotic proteins are more complex, as they: i) are on average longer [1–4] ii) are for a larger fraction multi-domain proteins [5–7], iii) contain more repeats domains [8] and iv) have more intrinsically disordered regions [9]. Several reasons for these differences have been proposed, including an increased need for regulation in the more complex eukaryotic cells.

The difference in average length between eukaryotic and prokaryotic proteins appears to be rather consistent over all phyla. Plant proteins are on average the shortes, while the proteins in single cellular eukaryota, fungi and alveolata, are as long as they are in metazoa. The length of eukaryotic proteins is obviously related to the presence of more domains, as multi-domain proteins are longer than single-domain proteins. Multi-domain proteins evolve by domain fusion events, primarily at N- or C-termini [10, 11]. Between these domains are often linker regions of various lengths. Multi-domain proteins have several potential advantages: the domains can act as independent functional units, providing multiple functions to a single protein, and the combination of domains can be essential for regulation. However, multi-domain proteins can be more difficult to fold correctly, and might be prone to aggregation. Therefore, it is possible that the compartmentalization of eukaryotic cells and a more elaborate chaperonin system are necessary requirements for some multi-domain proteins. The use of domains as building blocks also increases the repertoire of protein architectures (domain architectures) without the need to invent novel

domains. The number of unique domain architectures has increased significantly in the metazoan lineage [7] and it is possible that differences in the chaperonin systems have contributed to the increase in multi-domain eukaryotic proteins [12].

The increased number of repeats appears to be mainly a feature of multi-cellular organisms [8]. Repeat proteins are common in signalling and are associated with some cancers. These repeats have been proposed to provide the eukaryotes with an extra source of variability to compensate for low generation rates [13].

With a larger fraction of multi-domain proteins, it follows that eukaryotic proteins should have more linker regions - connecting the domains. Linker regions vary substantially in lengths, are longer in eukaryotes and are often intrinsically disordered [14]. It has been shown that variation of length between homologous proteins can largely be attributed to changes in the length of these regions [15].

The origin of the increased amount of intrinsically disordered regions in eukaryotic proteins is less well understood. Intrinsic disorder is frequent in all eukaryotic phyla, and even among viral proteins [16]. There is a spectrum of different types of disorder, spanning from increased flexibility to completely disordered proteins. Intrinsically disorder proteins exist among many classes of proteins with different functions, but the difference between eukaryotic and prokaryotic proteins appears to be maintained. On average less than 10% of the residues in prokaryotes are in disordered regions compared with 20% in eukaryotes [17].

It has been proposed that intrinsic disorder, as well as the other features separating eukaryotic and prokaryotic proteins, is a result of low selective pressure and small effective population size [17]. The authors argue that low selective pressure in eukaryotes causes the expansion of non-coding regions in eukaryotic genomes, as there is no strong purifying selection to keep it compact. In this model the increased disorder would be a "symptom" of the low pressure to maintain a compact genome. However, the large number of functionally important intrinsically disordered regions, see for instance a number of recent reviews [18–20], would argue against this.

One possible reason for this selective pressure is that post-translational modifications occur preferentially in intrinsically disordered regions [21]. Post-translational modification, in particular phosphorylation, is a fundamental mechanism in the regulation of eukaryotic cell differentiation, as well as many other processes [22].

Since the discovery of intrinsic disorder in proteins [17, 23] it has been clear that this property is a particular feature of eukaryotic proteins [9]. It should, however, be remembered that the vast majority of studies of intrinsic disorder are based on predictions [24]. Predictions of intrinsic disorder are based on frequency and patterns found in the amino acid sequences. Although the best predictors use factors such as conservation and correlation between amino acid positions, even simple predictors that are just based on the amino acid frequency can be used to detect the difference between eukaryotes and prokaryotes, i.e. there should be an underlying difference in the amino acid sequences that explains the difference in intrinsic disorder between eukaryotes and prokaryotes.

Polar and charged amino acids, together with proline, are the most disorder-promoting residues. Thus, proteins with a higher fraction of these types of residues are (predicted to be) more disordered, and vice-versa, disordered proteins should contain a higher frequency of these amino acids. One of the simplest of all predictors is the TOP-IDP scale [25]. TOP-IDP quantifies the "disorder propensity" for each amino acid. The average TOP-IDP value is significantly higher for eukaryotic proteins than for prokaryotic proteins, showing that there should be a consistent difference in amino acid distributions.

Amino acid frequency describes many features of a protein. It is also well known that for most positions in a sequence almost any amino acid is allowed. This is highlighted that by the fact that most protein domain families contain members that have less than 20% sequence identities and where each amino acid have been mutated on average up to five times [26].

Protein design experiments have also shown that it is often possible to design functional proteins with a limited, or biased, set of amino acids. This can be exemplified by some extreme cases such as the design of a protein without any charged residue [27]. This indicates that over evolutionary time there should be a rather large possibility for amino acid frequencies to change [28]. This means that if there is a pressure to change the amino

acid frequencies in the entire proteome, it should be possible for an organism to adapt to this.

The general trend of amino acid gains and lost have been studied before [29], and it was proposed that the amino acids that appeared to increase in frequency (except serine), were not incorporated in the first genetic code. However, the statistical methodology used in that study has been questioned [30].

It has also been reported that histidine and serine frequencies increase from high temperature thermophiles to prokaryotic mesophiles and further to eukaryotes [31]. Valine shows the opposite trend. It is also possible that a trend of increasing polar amino acids can be explained if the universal last common ancestor (LUCA) was oily [32], i.e it was enriched in hydrophobic amino acids. In this scenario it would be present a selective pressure to increase the number of polar amino acids, and thereby the predicted disorder. But why this increase only occurred in eukaryotes is not explained.

In this study we ask what are the molecular properties that determine the difference in intrinsic disorder between eukaryotes and prokaryotes. Firstly, we show that largely the difference in disorder can be contributed to that linker regions in eukaryotes are more disordered as well as more abundant than corresponding regions in prokaryotic proteins. Secondly, we show that the difference in disorder in these regions can largely be attributed to an increase of serine residues. Serine is one of the most disorder promoting residues and the fraction of serine is close to 8% of all residues in eukaryotes compared to less than 6% in prokaryotes. This difference is comparable to the difference in length between eukaryotic and prokaryotic proteins. A possible explanation for this increase is the use if serine as a target for regulatory phosphorylation in linker regions.

Material and Methods

Datasets

In this study we used two datasets. The first consists of all complete proteomes from Uniprot [33] as of December 2017. This dataset contains 36,781,033 sequences from 9288 genomes. The proteomes are divided into 506 viruses, 7320 Bacterial, 1053 Eukaryotic and 409 Archaeal ones. Length, disorder and other properties were analysed for the entire dataset, as described below. Here, all species from the following taxa, Mycoplasma, Spiroplasma, Ureaplasma and Mesoplasma were ignored as they have another codon usage - which influence the expected amino acid frequencies.

For the second dataset we started with the set of protein domain families from Pfam [34, 35] that are present in both bacterial and eukaryotic genomes. The smaller kingdoms Archaea and Viruses were ignored here. We retained all Pfam domains present in at least five bacterial and five eukaryotic species among the annotated "full" alignments in Pfam. This resulted in a set of 3950 Pfam domains. For each of these domains, we extracted from Uniprot the full-length sequences of the proteins containing them. This resulted in a set of 13,659,175 proteins, of which 4,932,465 eukaryotic and 8,726,710 bacterial.

Next, we divided each protein into regions, see Figure 1. The regions of each protein corresponding to any of the 3950 Pfam domains that exist both in prokaryotes and eukaryotes are referred to as "Shared domains". All regions assigned to any other Pfam domain are "Exclusive domains"; and everything that is not assigned to a Pfam domain is a classified as a "Linker" region. We analysed length, amino acid frequencies and other properties for the full-length proteins as well as each region independently.

Disorder prediction

For each protein, we estimated the intrinsic disorder by using two tools: IUPred [36] and the TOP-IDP scale [25]. IUPred exploits the idea that in disordered regions, amino acid residues form less energetically favourable contacts than the ones in ordered regions. IUPred does not rely on any external information besides the amino acid

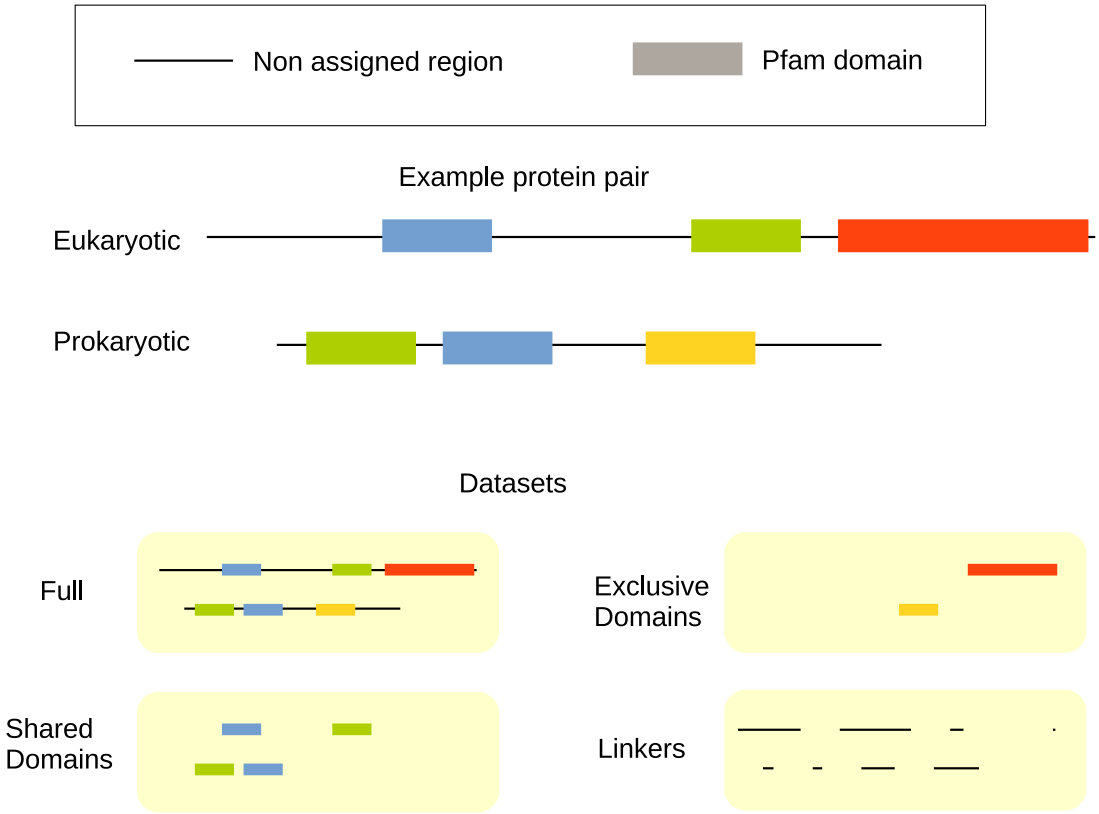


Figure 1. Representation of protein regions in the dataset of shared domains. Here a protein is divided into three potential regions: Shared Domains, Exclusive Domains and Linkers. By definition in this dataset all proteins contain at least one shared domain, but note that a protein can contain several domains of each type, i.e. it can contain more than one shared domain etc. Proteins without shared domains are ignored in this dataset.

sequence, and for this reason is extremely fast and suitable to predict disorder for a large dataset. We used IUPred in both its variants, "long" and "short". For each protein, we used the default cut-off and assigned a residue to be disordered if its IUPred value is greater than 0.5. The lengths of the regions were ignored here.

The TOP-IDP scale [25], assigns a disorder-propensity score to each amino acid, and it is based on statistics on previously published scales. For each protein, a TOP-IDP value is calculated as the average of the TOP-IDP values of all its residues.

Analysis

Properties, including amino acid type and disorder, of all residues were analysed independently. Comparisons were performed between all proteins in different kingdoms as well as for the second dataset between the four regions: full, shared domains, exclusive domains, linkers, see Figure 1.

It can be noted that even a small difference in a property between two differences is significant due to the large number of samples. The least significant difference we observe is for iupred_long that predicts 4.3% of the bacterial residues in the shared domains to be disordered, while 4.6% of the eukaryotic residues in these domains are predicted to be disordered. The P-value for this small difference is very significant ($< 10^{-8}$). All other P-values are smaller than 10^{-200}. Therefore, we do not believe it is of great relevance to report each individual P-value; what is more interesting is the magnitude of the differences. For this we use a two-sampled Z-test, see below.

Similarly, to compare differences in properties between kingdoms we have used the variation between species. Here, the Mann–Whitney U test [37] was used to evaluate if two distributions of e.g. amino-acid frequencies were similar or not. For instance all differences between bacteria and eukaryotes except five (frequencies of arg, lys, asp, glu and tyr) are very significant ($P < 10^{-11}$). However, the significance levels are strongly dependent on the number of species in each kingdom. This can for instance be seen by studying the average TOP-IDP scores of bacteria (0.064), archaea (0.072) and viruses (0.078). Here, the P-values for the difference between bacteria and archaea, i.e. the two kindoms with most samples, is $5 * 10^{-20}$ while the comparable in size difference between archaea and viruses is only $1 * 10^{-5}$. Therefore, we do not believe the statistical significance of each difference is very relevant. Instead we decided to also here test the sizes of the differences using the two-sampled Z-test.

To evaluate the magnitude of the difference between two sets we used a two-sample Z-test for comparing two means:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Here $\bar{X}_1$ and $\bar{X}_2$ represent the mean value of a property in the two kingdoms and σ_1^2 and σ_2^2 are the corresponding standard deviations. This comparison has the advantage that it is independent on sample size. The comparison of average TOP-IDP scores between bacteria, archaea and viruses provides comparable Z-scores, 0.36 vs 0.15. Further, as can be seen below this test results in comparable and biologically meaningful estimates for all the difference properties of the proteins that we study.

Results

In this result section, we first compare properties of proteins among the four kingdoms of life, eukaryota, bacteria, archaea and viruses. We use the list of 9288 complete proteomes, comprising more than 36 million protein sequences. Thereafter we go on to compare properties of different regions in eukaryotic and bacterial proteins. This is done by using the subset consisting of 14 million sequences that contain at least one of the 3950 Pfam domains that are common to eukaryota and bacteria.

Eukaryotic proteins are longer and more disordered.

As has been documented many times before, eukaryotic proteins are longer and more disordered than prokaryotic proteins, see Figure 2. Both bacterial and archaeal proteins have a median length of approximately 300 residues vs. about 400 for eukaryotic proteins. Some viruses contain very long protein sequences as they are coded as polyproteins [38], but the median length is similar to eukaryotic proteins.

Using all three different measures of intrinsic disorder, eukaryotic proteins are more disordered than prokaryotic ones, see Figure 2b-c. In agreement with earlier studies [17, 39–41] on average less than 10% of the residues in prokaryotes are in disordered regions compared with 20% in eukaryotes and about 12% in viral proteins.

All differences between the kingdoms are strongly significant (P$< 10^{-100}$), but also dependent on the number of genomes in each kingdom. Therefore we do think it is more relevant and meaningful to use the two-sample z-test to compare the different kingdoms. Length and disorder are between one to two standard deviations higher

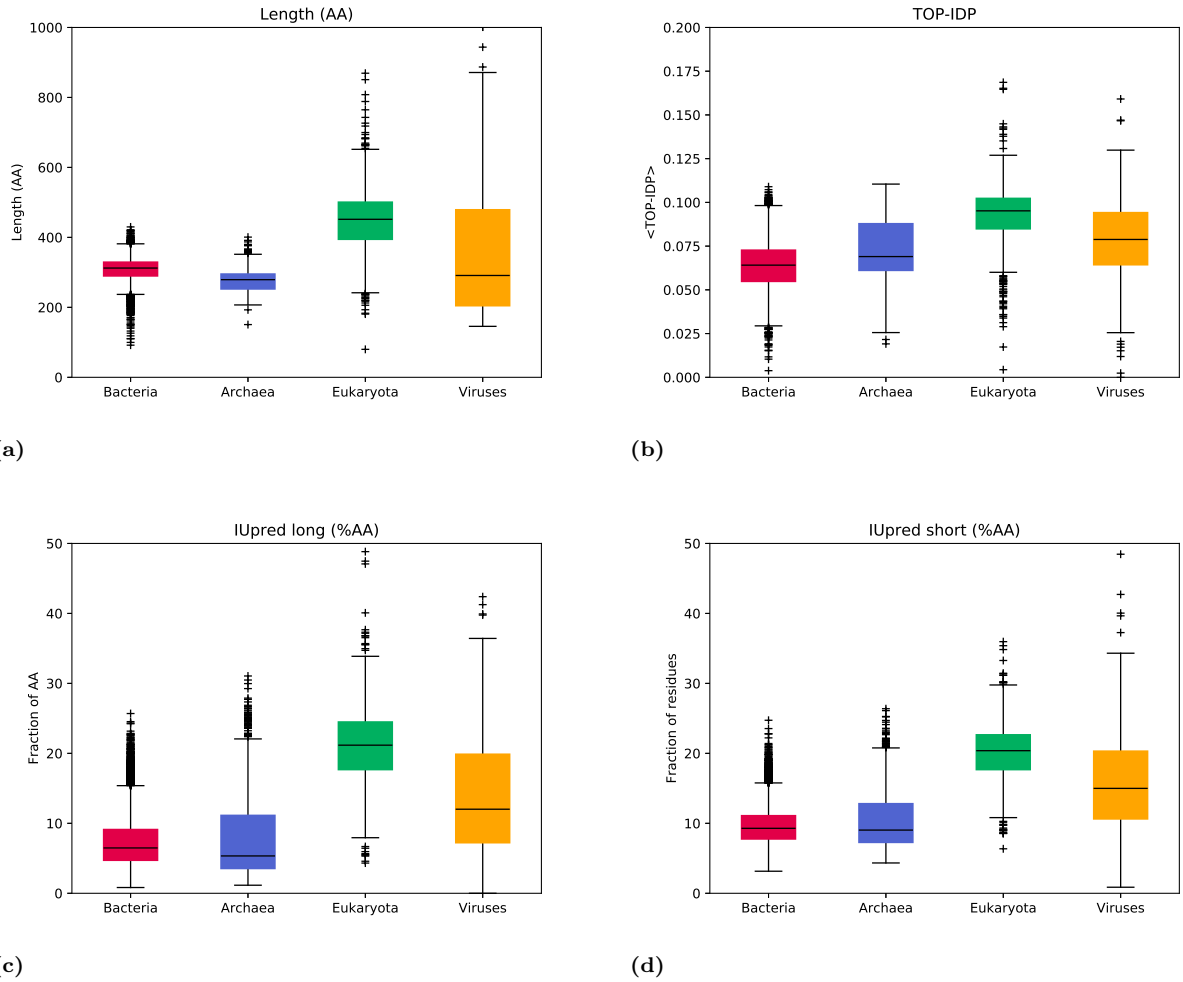


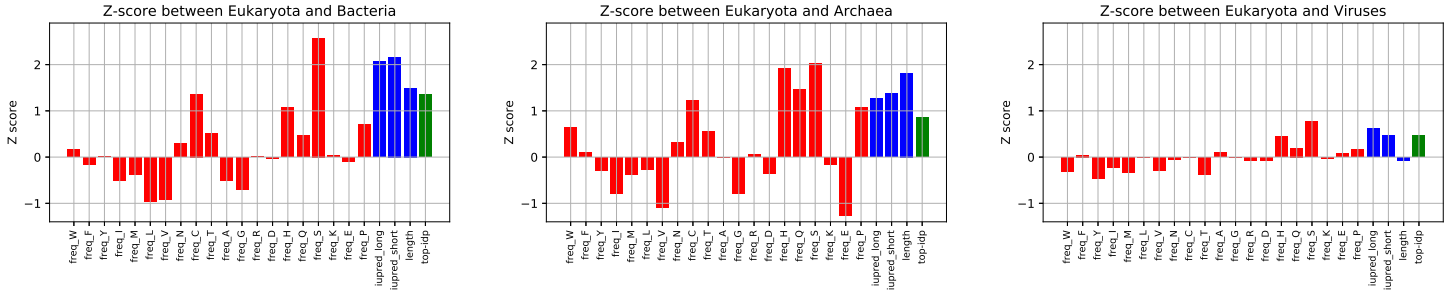
Figure 2. Average properties of proteins from different kingdoms; (a) average length, (b) average TOP-IDP; scores and fraction of residues predicted to be disordered by (c) IUPred-short and (d) IUPred-long

in eukaryotes than in either of the two prokaryotic kingdoms, see blue and green bars in Figure 3. The differences between eukaryotic and viral proteins and between the prokaryotes are much smaller, see also Figure S1.

Systematic amino acid bias between eukaryotes and prokaryotes.

Intrinsic disorder, as studied in this and many other papers, is a feature predicted from the amino acid sequence of a protein. Therefore, the differences in intrinsic disorder content observed between the kingdoms should have its origin in differences in amino acid frequencies.

We studied the average frequency of each amino acid in each kingdom. It can be noted that even small difference such as the difference in trp frequency between eukaryotes and bacteria (12.7% vs 12.1%) is highly



(a) **(b)** **(c)**

Figure 3. Two-sided z-score of properties differences between eukaryotes and the other kingdoms. Positive numbers represent an overrepresentation in eukaryotes. Amino acid frequencies are shown in red, intrinsic disorder in blue and length in green.

significant ($P < 10^{-11}$ using the Mann Whitney U test [37]). The P-value for differences in serine frequencies is $P < 10^{-300}$. Therefore, to compare the size of the differences we used the two-sample Z-test for each amino acid frequency, see red bars in Figure 3.

There appear to be systematic differences in amino acid frequencies between eukaryotic and prokaryotic proteins. Both in respect to bacteria and archaea, certain amino acids are more or less frequent. Serine, cysteine and histidine are clearly overrepresented in eukaryotes and to a lesser extent asparagine, threonine and glutamine. The differences between viruses and eukaryotes are much smaller, but follow mostly a similar trend. The compensatory underrepresentation of amino acids in eukaryotes seems to be spread over a set of smaller amino acids, including ala, val and gly. None of these underrepresentations are larger than one standard deviation suing the Z-test.

In pure numbers the largest difference between eukaryotic and bacterial proteins can be seen in alanine and serine frequencies. In eukaryotes 7.9% of the amino acids is serine, and 7.6% are alanine, while in bacteria ala is almost twice as abundant than ser, 10.3% vs. 5.7%.

The Z-test difference in serine frequency is as large as the differences observed for length or disorder, i.e. it would be equally correct to claim that eukaryotic proteins are enriched in serine as claiming that eukaryotic proteins are longer or more disordered than prokaryotic proteins.

Eukaryotic proteins are longer because they have longer linker regions.

Eukaryotic proteins are longer than prokaryotic proteins, see Figure 2. This is largely due to that multi-domain proteins are almost twice as frequent in eukaryotes; further long repeat proteins are also more abundant [4]. To examine the effect of this we created a dataset of proteins that contain at least one Pfam domain that exists both in eukaryotes and in bacteria. The proteins are then divided into three regions: shared domains, exclusive domains and linkers, see Figure 1.

As it is well known, eukaryotic proteins are on average longer than bacterial proteins [42], and this is also observed for the 14 million proteins with shared domains, see Figure 4a. The number of residues in the shared domains is roughly equal, slightly more than 200 residues per protein on average. Here, it should be remembered that the length of the shared regions might correspond to one or several domains in a protein and that this dataset does only contain proteins with at least one shared domain.

Next, it can be seen that the other regions are longer in eukaryotic proteins. The average number of residues

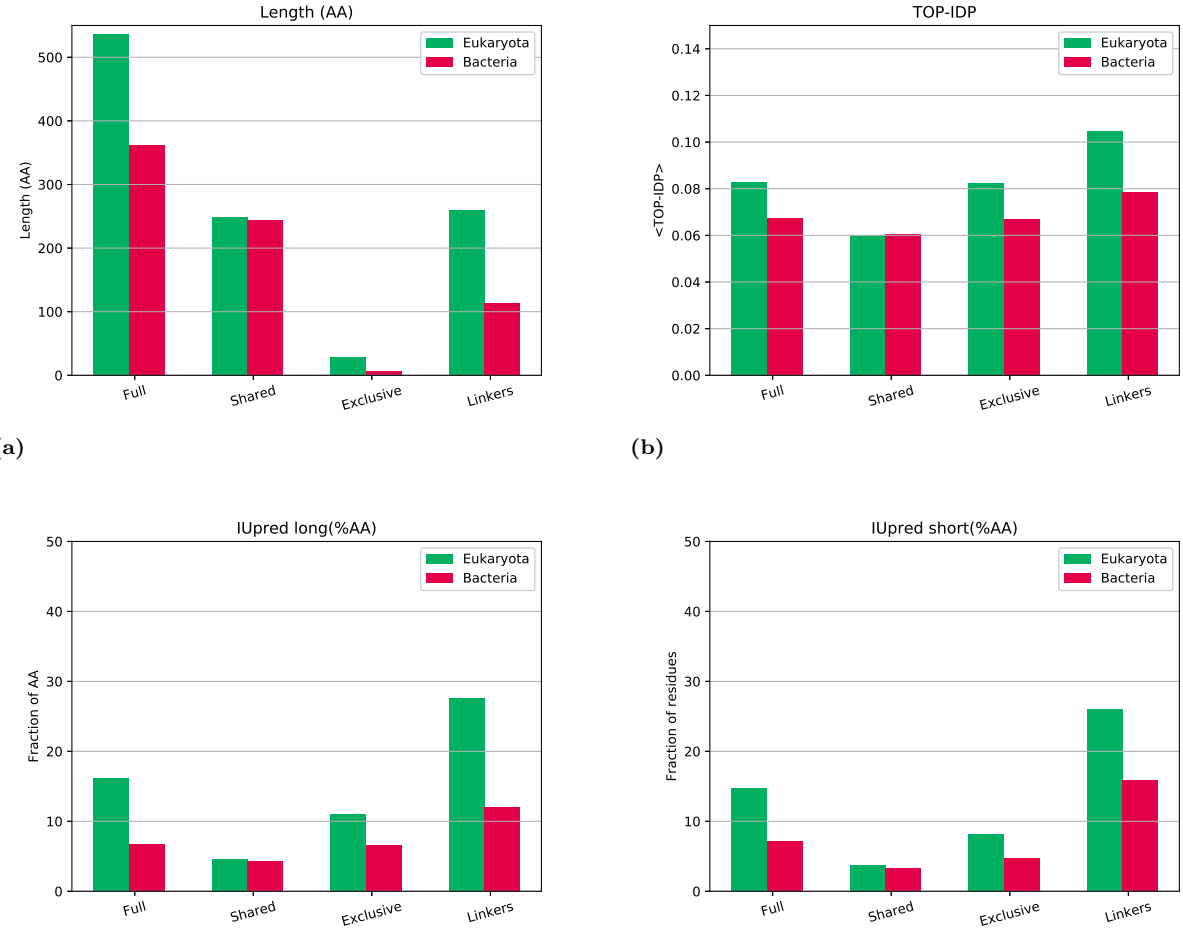


Figure 4. For eukaryotes (red) and bacteria (green), it is shown length (a) and intrinsic disorder in different parts of proteins, estimated with the TOP-IDP scale (b), IUPred-long (c) and IUPred-short (d).

assigned to kingdom-specific domains is lower in bacteria than eukaryotes, but the overall number of residues assigned to these unique domains is small, as only a small fraction of all proteins contain a kingdom-specific domain.

The largest contribution to the length difference is due to linkers. In eukaryotes about half of the residues are assigned to linkers, while in bacteria less than a third. The difference in average length can therefore mainly be explained by eukaryotic proteins having longer linker regions [43]. It should be remembered that these regions are not necessarily only linkers, but they can also contain unassigned domains and elements of N- or C-terminal extensions of existing domain families [44, 45]. Further, this dataset is not perfectly representing all proteins, as protein that only contain kingdom-specific domains are ignored. This include many of the long repeat containing proteins.

	Eukaryota				Bacteria			
	Full	Shared domains	Exclusive domains	Linkers	Full	Shared domains	Exclusive domains	Linkers
TRP	1.33%	1.51%	1.35%	1.15%	1.23%	1.22%	1.35%	1.26%
PHE	4.04%	4.53%	4.35%	3.54%	3.81%	3.93%	3.96%	3.57%
TYR	3.00%	3.37%	3.36%	2.61%	2.85%	2.9%	3.20%	2.72%
ILE	5.33%	6.05%	5.29%	4.64%	5.83%	6.09%	5.83%	5.29%
MET	2.26%	2.27%	2.24%	2.25%	2.37%	2.25%	2.02%	2.66%
LEU	9.40%	9.82%	9.96%	8.94%	10.14%	10.22%	10.57%	9.94%
VAL	6.53%	7.22%	6.30%	5.90%	7.41%	7.74%	7.33%	6.71%
ASN	4.20%	4.09%	4.35%	4.28%	3.44%	3.38%	3.63%	3.55%
CYS	1.80%	1.86%	2.59%	1.64%	0.91%	0.97%	0.86%	0.78%
THR	5.55%	5.48%	5.26%	5.65%	5.48%	5.41%	5.41%	5.64%
ALA	7.59%	7.70%	6.72%	7.58%	10.33%	10.36%	9.58%	10.31%
GLY	6.74%	7.42%	5.58%	6.21%	7.95%	8.37%	7.05%	7.10%
ARG	5.50%	5.11%	5.68%	5.86%	5.98%	5.78%	6.23%	6.38%
ASP	5.44%	5.34%	5.46%	5.53%	5.63%	5.62%	5.79%	5.65%
HIS	2.47%	2.53%	2.48%	2.41%	2.11%	2.18%	1.97%	1.95%
GLN	3.93%	3.46%	4.36%	4.33%	3.55%	3.39%	4.00%	3.86%
SER	7.92%	6.84%	7.17%	9.04%	5.71%	5.51%	5.87%	6.13%
LYS	5.51%	5.26%	6.12%	5.67%	4.48%	4.30%	4.60%	4.87%
GLU	6.20%	5.67%	6.90%	6.64%	6.16%	6.00%	6.39%	6.47%
PRO	5.24%	4.45%	4.47%	6.08%	4.63%	4.39%	4.35%	5.17%

Table 1. Amino acid frequencies for each region and kingdom. The amino acids are sorted according to the TOP-IDP scale.

Eukaryotic linkers are more disordered

Eukaryotic proteins are on average more disordered than bacterial ones. This is independent of the method used to evaluate disorder (TOP-IDP, IUPred-long or IUPred-short). Linker regions are more disordered than domains [24]. Therefore, the difference in disorder could be explained by linker regions being more abundant in eukaryotic proteins as shown above. However, in Figure 4 it can also be seen that linker regions in addition to being longer also are more disordered in eukaryotes. Both the length of linker regions and the amount of disordered residues in these regions contribute to the difference.

Interestingly, the presumably ancestral shared domains are less disordered regions than the kingdom specific domains in both kingdoms. Further, these domains are equally disordered in both kingdoms indicating that there is not a general trend to make things more disordered in eukaryotes. However, the unique domains in eukaryotes, in addition to being longer, also are more disordered.

Amino acid differences in different regions

Next, we calculated the amino acid frequency for each region and kingdom and compared them to each other, see Table 1 and Figures 5 and S2. Here, the amino acids are sorted by their TOP-IDP values, i.e. their disorder-promoting propensity. In both bacteria and eukaryotes the linker regions contain more of several disorder-promoting residues (ser, lys, glu, gln and pro) and less of the order promoting (hydrophobic) residues

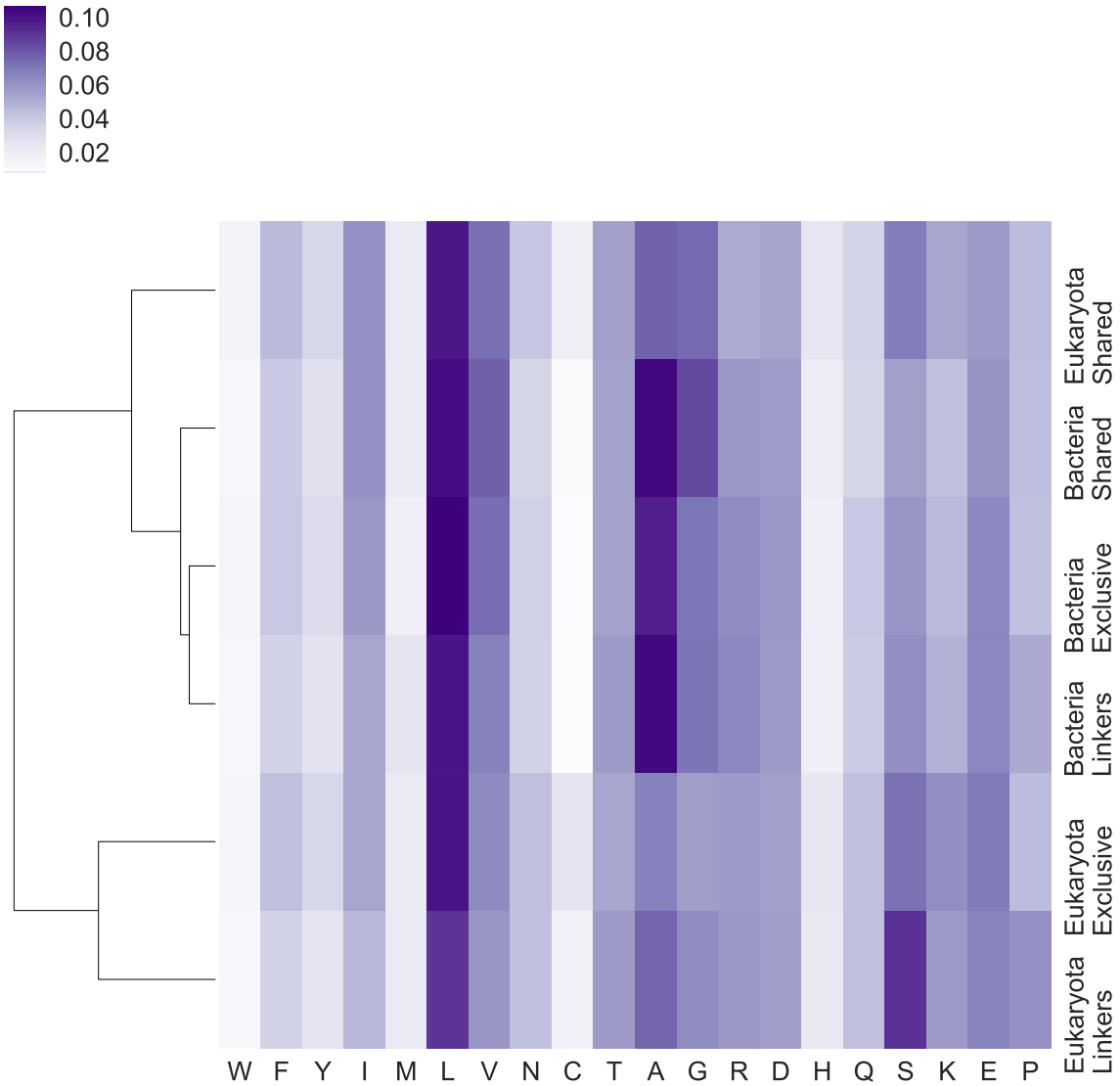


Figure 5. Distribution of amino acid frequencies in the regions, clustered according to their frequency profiles. The color of each cell represents the frequency of an amino acid in a dataset, according to the reference color bar.

(phe, tyr, ile, leu and val). It can also be noted that the unique domains in eukaryotes contain more cysteine, while the shared domains in both kingdoms are enriched in glycine.

In Figure 5 the frequency of all amino acids in each region is shown. The amino acids are sorted by TOP-IDP with the most disorder-promoting residues to the right. A few differences stand out. All bacterial regions are enriched in alanine. On average 10.3% of the bacterial amino acids is alanine vs. 7.6% in eukaryotes. Another outlier is serine, which makes up 9.0% of the eukaryotic linker regions vs. $\approx 7\%$ in other eukaryotic regions and only 5.7% in bacteria, see Table 1. Other differences include that cys is more frequent in eukaryotic proteins, while

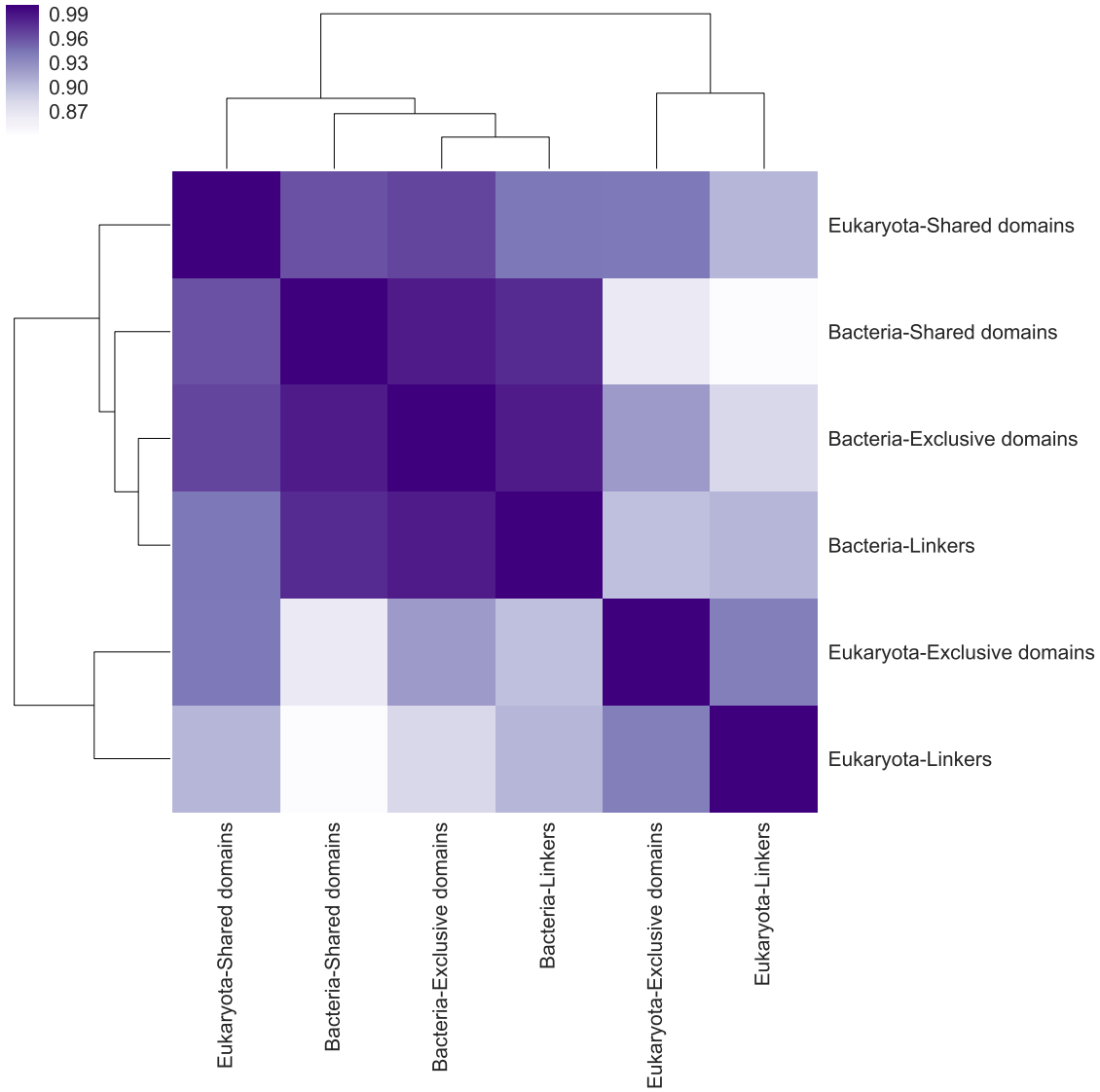


Figure 6. Heat map showing the similarity of amino acid frequency profiles in different regions as measured by the Pearson correlation coefficient. The color of each cell represents the Pearson correlation, according to the reference color bar.

gly is preferred in the bacterial exclusive domains.

Bacterial regions are similar to each other.

Figure 6 shows a heat map based on the Pearson correlations of the amino acid frequencies in each region. The amino acid distributions of the bacterial regions are very similar ($CC > 0.98$), while the variation between

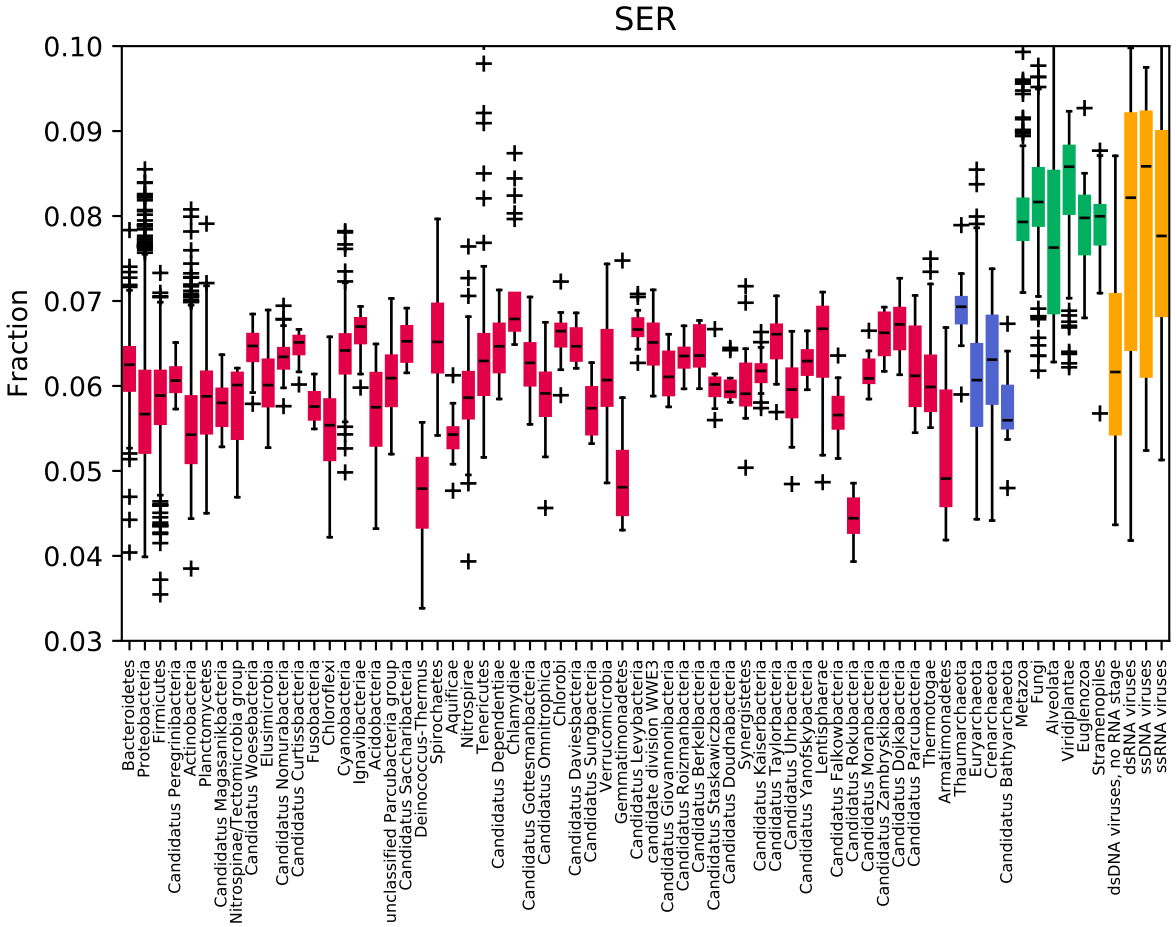


Figure 7. Frequency of serine in complete proteomes grouped by phylum. Bacterial groups are red, eukaryotic green, archaeal blue, and viral are yellow.

eukaryotic regions is much larger ($CC = 0.90 - 0.94$). Also, the eukaryotic shared domains are actually more similar to bacterial regions ($CC = 0.94 - 0.96$) than to the other eukaryotic regions ($CC = 0.90 - 0.94$). The amino acid frequencies of the eukaryotic unique domains and eukaryotic linker regions are also rather similar ($CC = 0.97$). This might indicate that the observed amino acid differences in eukaryotes is largely due to expansions of eukaryotic specific domains and linker regions. However, even among the shared domains the amino acid preferences have shifted somewhat in the same direction.

Discussion

Above we show that the difference in intrinsic disorder between prokaryotes and eukaryotes can mainly be attributed to two factors: eukaryotic proteins contain longer linker regions, and these linker regions are more disordered. To explain what makes them more disordered, we notice that there is a consistent difference in amino

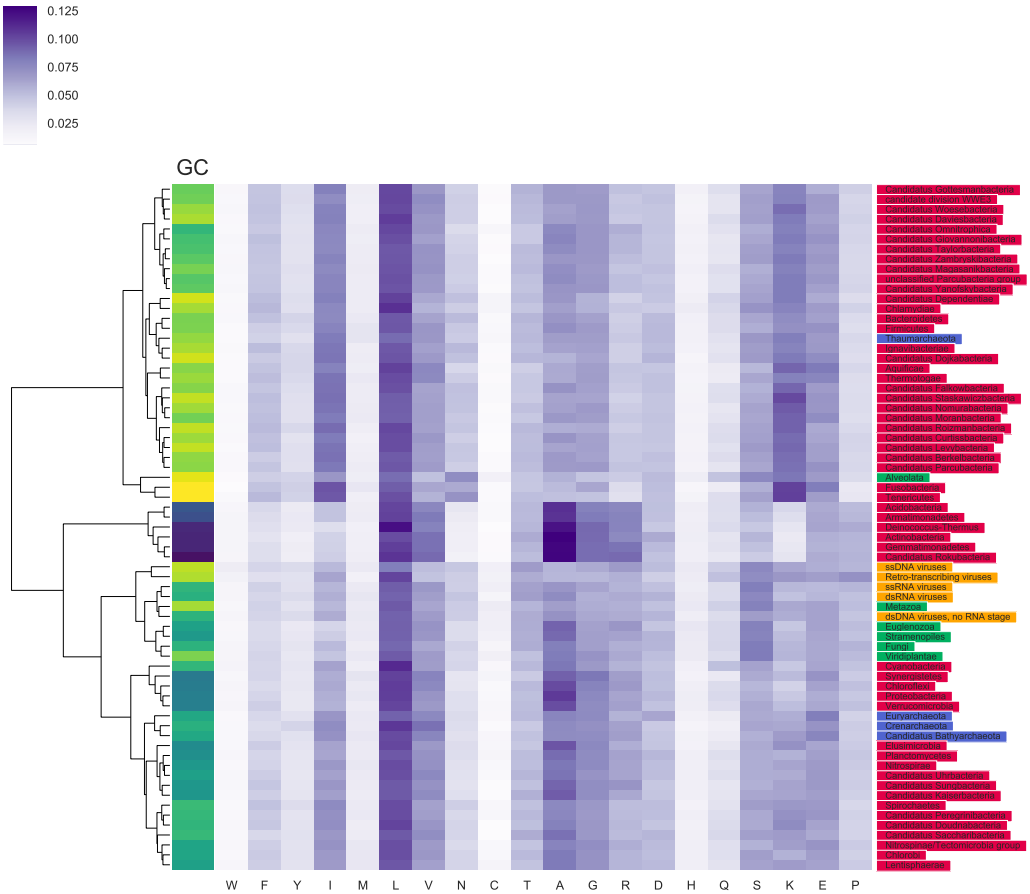


Figure 8. Heat map showing amino acid frequencies in different phyla. Phyla are clustered according to their frequency profiles. The color of each cell represents the frequency of an amino acid in a dataset, according to the reference color bar.

acid preference between bacteria and eukaryotes. In particular serine is much more abundant in eukaryotic proteins.

To analyse this in more detail we calculated the amino acid frequency of each completely sequenced proteome in the uniprot reference set. Thereafter, we grouped them by phylum and compared them, see Figure 7 and Supplementary Figures S4-S23. What can be observed is that serine stands out. Serine is consistently more frequent in all groups of eukaryotes than in any group of archaea or bacteria. For the other 19 amino acids the trends are less clear with at least one phylum crossing the prokaryotic-eukaryotic border. It is beyond the goals of this study to go through all amino acids.

Should we actually skip this ? It start diverging (and might be better suited for the.

In Figure 8 all phyla are clustered based on their amino acid frequencies. The names are coloured according to the kingdom and the bar to the left represents the average GC content of each phylum genome. All phyla can be divided into five groups (from the top): (i) low GC prokaryotes, (ii) extreme low GC genomes, (iii) extremely high GC prokaryotes, (iv) eukaryotes and viruses and (v) intermediate GC prokaryotes.

This clustering can be explained by two trends, (a) GC content and (b) eukaryotic-specific amino acid

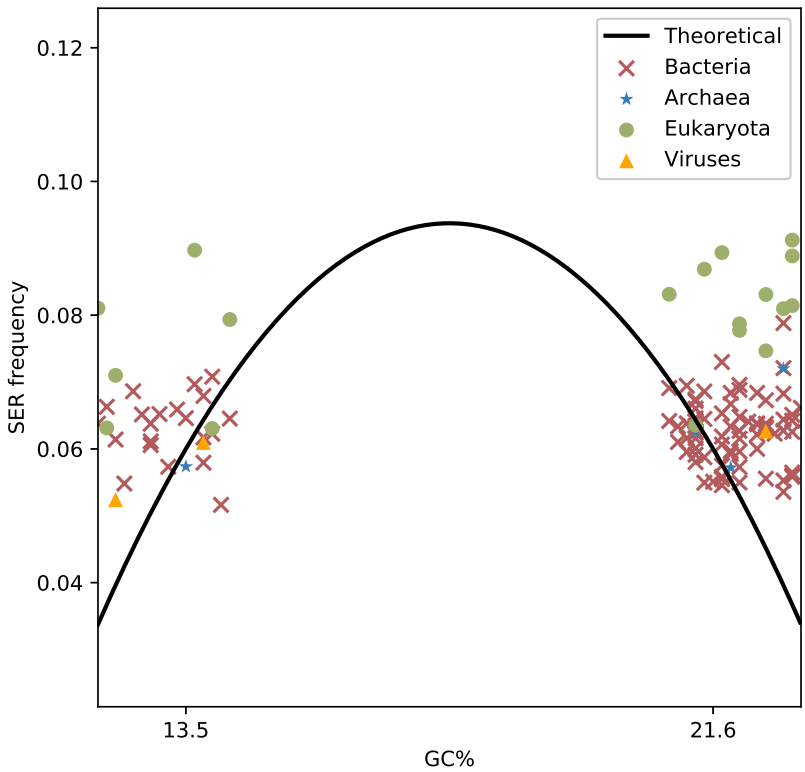


Figure 9. Average frequency of serine vs GC content of a genome for 9288 genomes. Genomes are colored by the kingdom they belong to using the same color scale as in other plots (see inset). A theoretical curve from the codon frequency as a function of GC is shown in black.

distributions. The effect of GC can for instance be observed in lys/arg frequencies. In the low GC group lysine is much more frequent than arg, while the reverse is observed in the high GC group. The high GC group is also enriched in alanine. Given the fraction of GC in the codons coding for ala (83%), arg (72%) and lys (17%) these shifts are easily understood.

With one exception all eukaryotes and viruses are found in a single cluster. The only eukaryotic phylum that is not included is alveolata, which clusters with two extremely GC-poor bacterial phyla. It has been reported that the ancestor of all plasmodium was extremely GC poor [46].

As already mentioned above, it is clear that serine frequencies are enriched in all the eukaryotic/virus phyla.

Serine

Why are eukaryotes enriched in serine? And when did it occur? First we examined if the GC content of an organism could be a factor. In Figure 9 it can be seen that the serine frequency is increased in eukaryotes

independently of the GC level. At any GC-level the eukaryotic proteins have more serines. About half of the viruses do also have as many serines as the eukaryotes. Other amino acids are strongly dependent on GC, see Supplementary Figure S24. In general amino acids coded by low GC codons are more frequent in low GC genomes but interesting differences exist that are beyond the goals of this study.

We examined the amount of serine in the prokaryotic phylum that has been proposed [47] to bridge the gap between prokaryotes and eukaryotes: the archaea Lokiarchaeota contain 6.4% serines, while one of the most primitive eukaryotes, *Giardia lamblia* contains 9.2%. Serine/threonine kinases are much more prevalent in eukaryotes, but also exist in bacteria [48]. For instance it has been reported to exist in Planctomycetes bacteria, but in the only fully sequenced genome of this phylum (*Planctomycetes bacterium GWA2_40_7*) there are only 6.1% serines in the 703 proteins. Further, the major family of ser/thr kinases, PFam family Stk19 (PF10494), only exists in eukaryotes and in Halanaerobiales. Among the 2783 Halanaerobium sequences in UniProt [33] there are 5.8% serines, typical of a prokaryote.

Some bacteria in the Planctomycetes, Verrucomicrobiae, and Chlamydiae bacterial superphylum have quite complex membranes [49]. However, all these phyla have typical Serine levels for bacteria.

One possible reason for the higher fraction of serine in eukaryotic organisms is that serine, together with threonine, is target for ser/thr kinases [50]. Phosphorylation of serine and threonine is one the most important regulatory pathways in eukaryotes, but also present in archaea [51]. We observe an increase only in ser, and not in thr or tyr, the other targets for kinases. This might be due to that about 75% of the known targets for kinases are serines [52].

Taking all this together indicates that the increase of serine is something that occurred early after LUCA [53].

It is also established that phosphorylation occurs frequently in intrinsically disordered sites [54]. This leads to a question: are eukaryotic proteins enriched in serine because they are disordered, or are they disordered because they are enriched in serine? When excluding serine the average TOP-IDP score of eukaryotic proteins is still higher than for bacterial proteins. When excluding sering the shared domains the bacterial proteins appears more disordered than the eukaryotic ones.

Conclusion

We show that in addition to the two well known distinct features that separate eukaryotic and prokaryotic proteins (length and disorder content), there are differences in amino acid frequencies clearly distinguishing the proteins. These differences are of a similar order of magnitude. Here, we focus on the amino acid with the largest difference, serine. Serine is much more frequent in eukaryotes than in prokaryotes, 8.1% vs 5.9%.

We show that in all regions of a protein serine is more frequent in eukaryotic than in bacterial proteins. Serine is a strongly disorder-promoting residue and is most frequent in eukaryotic linker regions. It is not unlikely that the necessity for regulatory mechanisms through phosphorylation of serines in predominantly disordered linker regions has been a driving force for the increased intrinsic disorder in eukaryotic proteins.

References

1. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 2001;310(2):311–325.

2. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. Protein Sci. 1998;7:445–456.

3. Liu J, Rost B. CHOP proteins into structural domain-like fragments. PROTEINS: Structure, Function and Bioinformatics. 2004;55:678–688.

4. Ekman D, Bjorklund AK, Frey-Skott J, Elofsson A. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 2005 Apr;348(1):231–243.

5. Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des. 1998;3(6):497–512.

6. Apic G, Gough J, Teichmann SA. An insight into domain combinations. Bioinformatics. 2001;17(Suppl 1):S83–89.

7. Ekman D, Bjorklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 2007 Oct;372(5):1337–1348.

8. Bjorklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Comput Biol. 2006 Aug;2(8):e114.

9. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn. 2012;30(2):137–149.

10. Weiner Jr, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. FEBS. 2006;273(9):2037–2047.

11. Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. Trends Biochem Sci. 2008 Sep;33(9):444–451.

12. Jacob E, Horovitz A, Unger R. Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study. Bioinformatics. 2007 Jul;23(13):i240–8.

13. Marcotte E, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol. 1999 Nov 15;293(1):151–160.

14. Wang M, Kurland CG, Caetano-Anolles G. Reductive evolution of proteomes and protein structures. Proc Natl Acad Sci U S A. 2011 Jul;108(29):11954–11958.

15. Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol. 2013 Dec;30(12):2645–2653.

16. Uversky VN. Intrinsic disorder here, there, and everywhere, and nowhere to escape from it. Cell Mol Life Sci. 2017 Sep;74(17):3065–3067.

17. Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. Evolution of intrinsic disorder in eukaryotic proteins. Cell Mol Life Sci. 2017 Sep;74(17):3163–3174.

18. Tompa P, Schad E, Tantos A, Kalmar L. Intrinsically disordered proteins: emerging interaction specialists. Curr Opin Struct Biol. 2015 Dec;35:49–59.

19. Pancsa R, Tompa P. Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. Trends Biochem Sci. 2016 Nov;41(11):898–906.

20. Pauwels K, Lebrun P, Tompa P. To be disordered or not to be disordered: is that still a question for proteins in the cell? Cell Mol Life Sci. 2017 Sep;74(17):3185–3204.

21. Pejaver V, Hsu WL, Xin F, Dunker AK, Uversky VN, Radivojac P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. Protein Sci. 2014 Aug;23(8):1077–1093.

22. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32(3):1037–1049.

23. Ahrens J, Dos Santos HG, Siltberg-Liberles J. The Nuanced Interplay of Intrinsic Disorder and other Structural Properties Driving Protein Evolution. Mol Biol Evol. 2016 May;.

24. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci. 2017 Sep;74(17):3069–3090.

25. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15(9):956–63. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/18991772`.

26. Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins. 2009 Nov;77(3):499–508.

27. Kurnik M, Hedberg L, Danielsson J, Oliveberg M. Folding without charges. Proc Natl Acad Sci U S A. 2012 Apr;109(15):5705–5710.

28. Singer GA, Hickey DA. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol. 2000 Nov;17(11):1581–1588.

29. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, et al. A universal trend of amino acid gain and loss in protein evolution. Nature. 2005 Feb;433(7026):633–638.

30. Goldstein RA, Pollock DD. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. Mol Biol Evol. 2006 Jul;23(7):1444–1449.

31. Tekaia F, Yeramian E. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics. 2006 Dec;7:307.

32. Mannige RV, Brooks CL, Shakhnovich EI. A universal trend among proteomes indicates an oily last common ancestor. PLoS Comput Biol. 2012;8(12):e1002839.

33. Consortium TU. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D142–8. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/19843607`.

34. Sammut SJ, Finn RD, Bateman A. Pfam 10 years on: 10 000 families and still growing. Brief Bioinform. 2008;9:210–219. Available from: `http://dx.doi.org/10.1093/bib/bbn010`.

35. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016 Jan;44(D1):D279–85.

36. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005 Apr;347(4):827–39. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/15769473`.

37. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics. 1947;18(1):50–60.

38. Yost SA, Marcotrigiano J. Viral precursor polyproteins: keys of regulation from replication to maturation. Curr Opin Virol. 2013 Apr;3(2):137–142.

39. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002 Oct;27(10):527–33. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/12368089`.

40. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. Biochem Cell Biol. 2010;88:167–174.

41. Ekman D, Elofsson A. Identifying and quantifying orphan protein sequences in fungi. J Mol Biol. 2010 Feb;396(2):396–405.

42. Gerstein M. Integrative database analysis in structural genomics. Nat Struct Biol. 2000;7(suppl):960–963.

43. Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A. Domain rearrangements in protein evolution. J Mol Biol. 2005 Nov;353(4):911–923.

44. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA. Structural diversity of domain superfamilies in the CATH database. J Mol Biol. 2006 Jul;360(3):725–741.

45. Mistry J, Coggill P, Eberhardt RY, Deiana A, Giansanti A, Finn RD, et al. The challenge of increasing Pfam coverage of the human proteome. Database (Oxford). 2013;2013:bat023.

46. Nikbakht H, Xia X, Hickey DA. The evolution of genomic GC content undergoes a rapid reversal within the genus Plasmodium. Genome. 2014 Sep;57(9):507–511.

47. Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015 May;521(7551):173–179.

48. Pereira SF, Goss L, Dworkin J. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. Microbiol Mol Biol Rev. 2011 Mar;75(1):192–212.

49. Santarella-Mellwig R, Pruggnaller S, Roos N, Mattaj IW, Devos DP. Three-dimensional reconstruction of bacteria with a complex endomembrane system. PLoS Biol. 2013;11(5):e1001565.

50. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res. 1998 Oct;8(10):1038–1047.

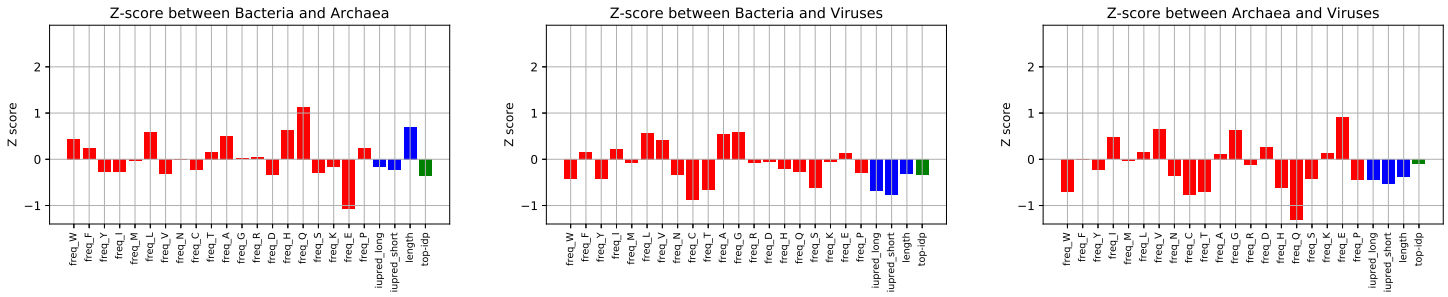
51. Kennelly PJ. Protein Ser/Thr/Tyr phosphorylation in the Archaea. J Biol Chem. 2014 Apr;289(14):9480–9487.

52. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho.ELM: a database of phosphorylation sites–update 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D261–7.

53. Forterre P, Philippe H. The last universal common ancestor (LUCA), simple or complex? Biol Bull. 1999 Jun;196(3):373–5; discussion 375–7.

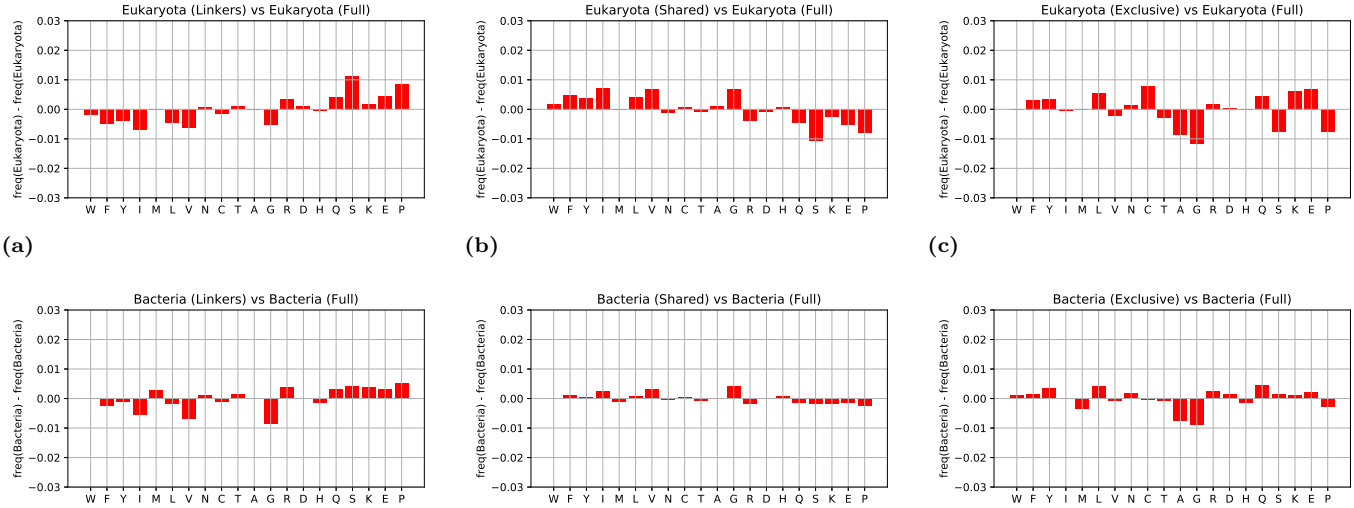
54. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999 Dec;294(5):1351–1362.

Supporting Information



(a) **(b)** **(c)**

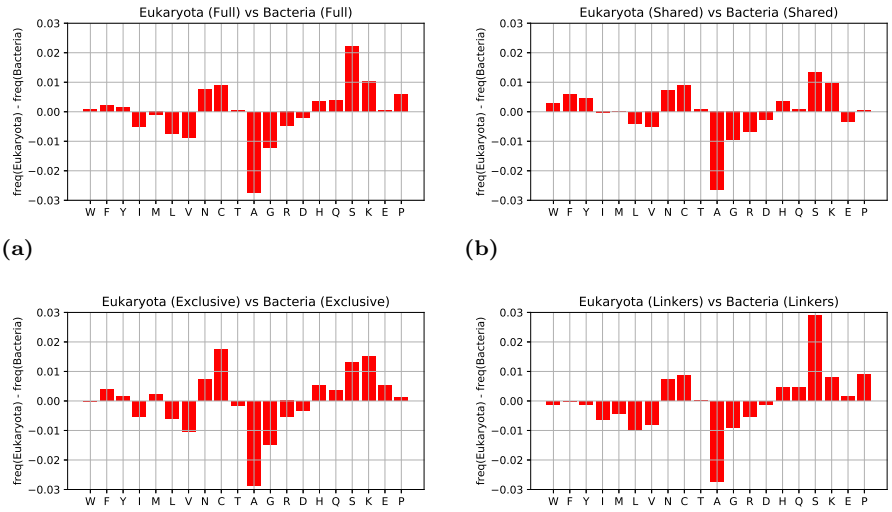
Figure S1. Z-score of comparisons of of properties between kingdoms.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure S2. Difference in amino acid frequencies between different regions in eukaryota (a-c) and bacteria (d-e)



(a) **(b)** **(c)** **(d)**

Figure S3. Difference in amino acid frequencies between Eukaryota and bacteria in different datasets

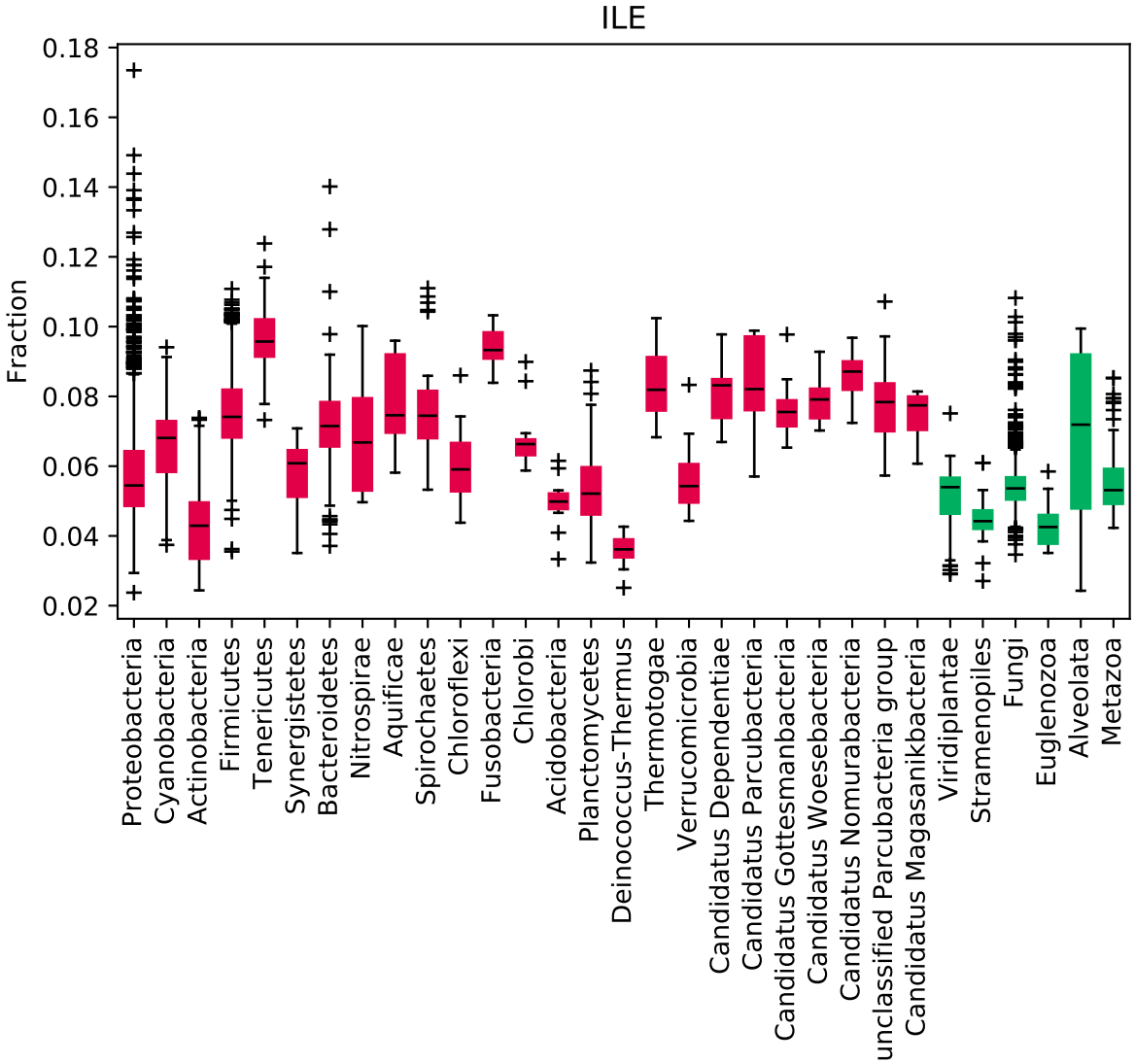


Figure S4. Frequency of isoleucine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

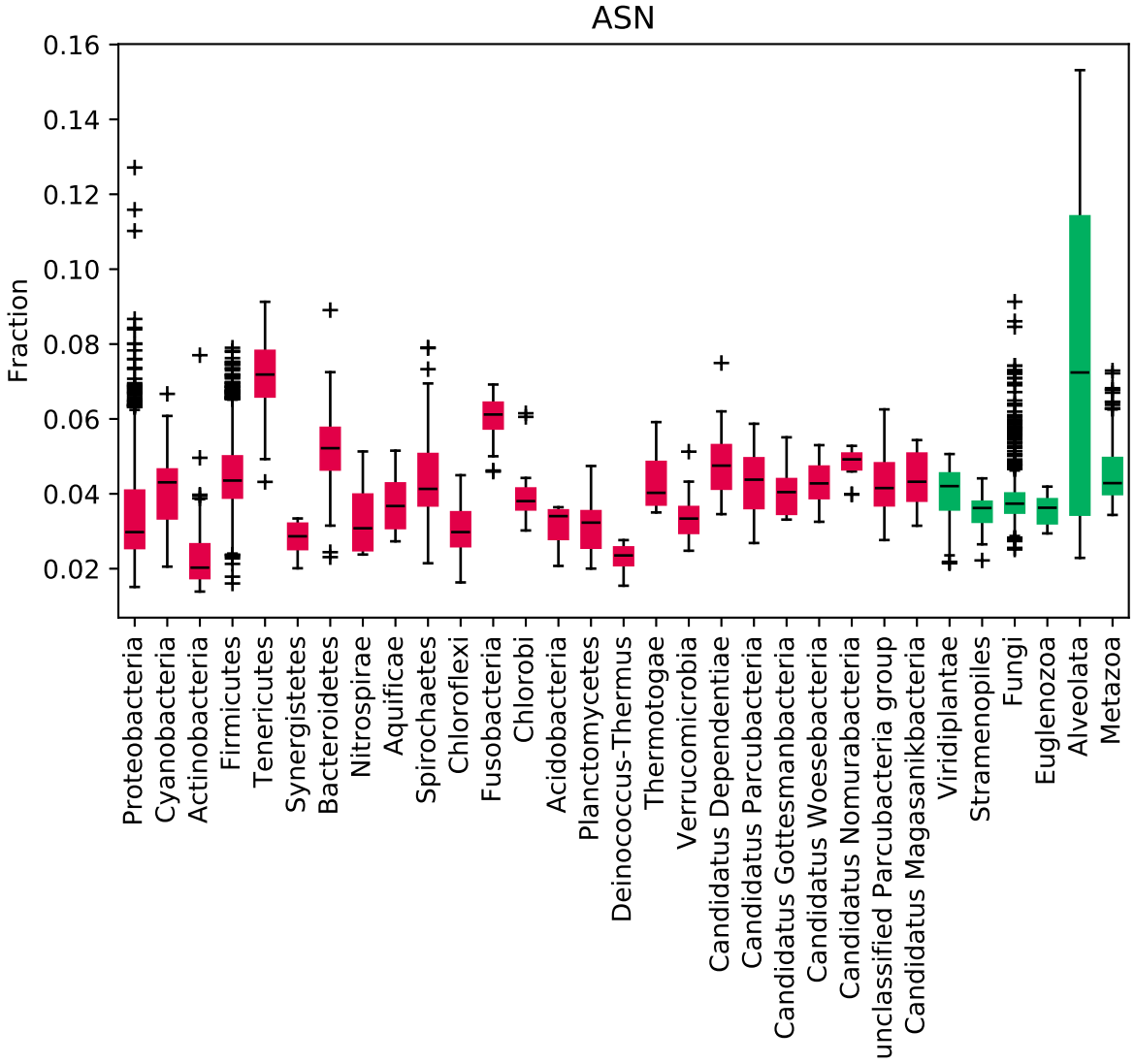


Figure S5. Frequency of asparagine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

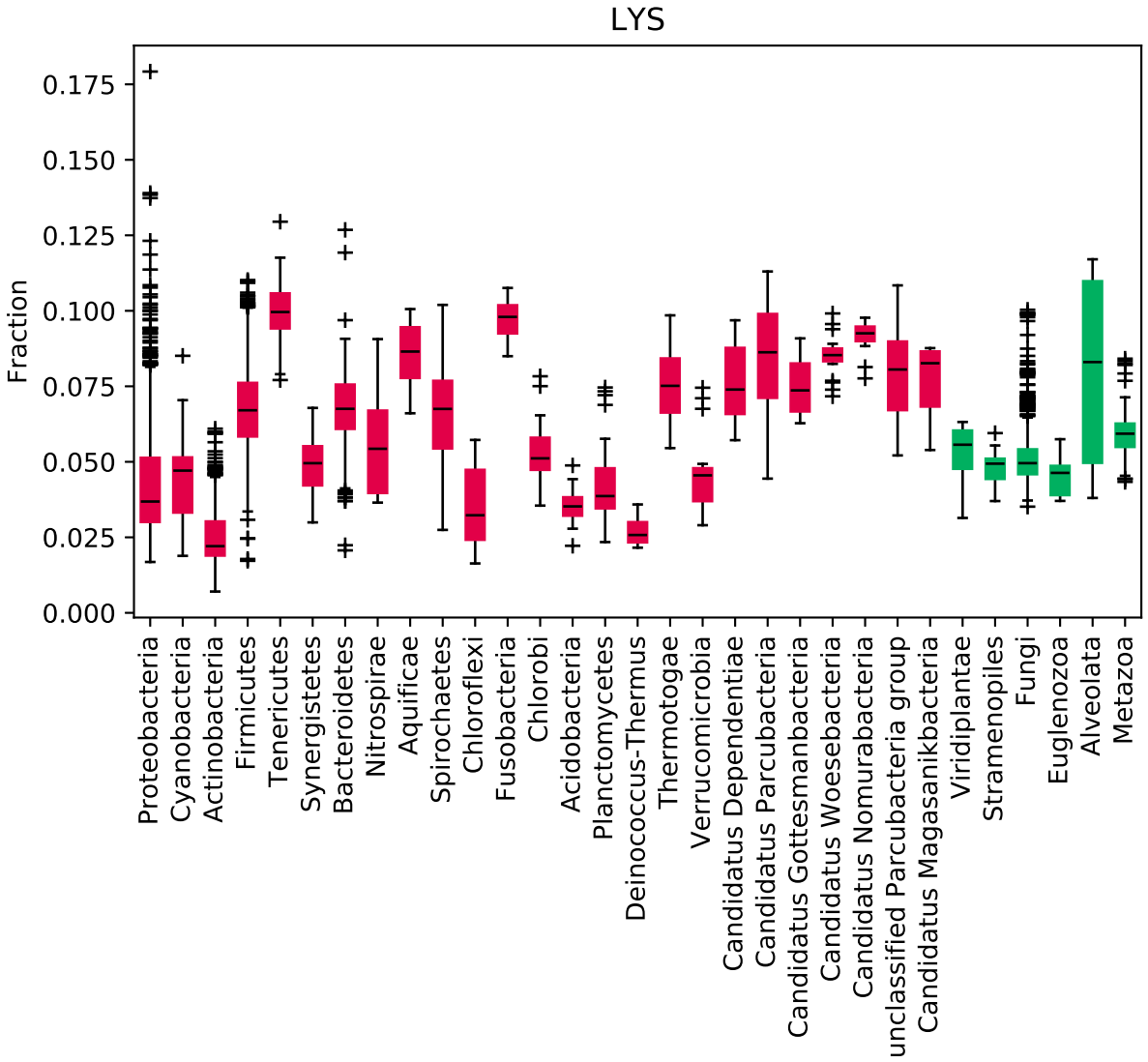


Figure S6. Frequency of lysine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

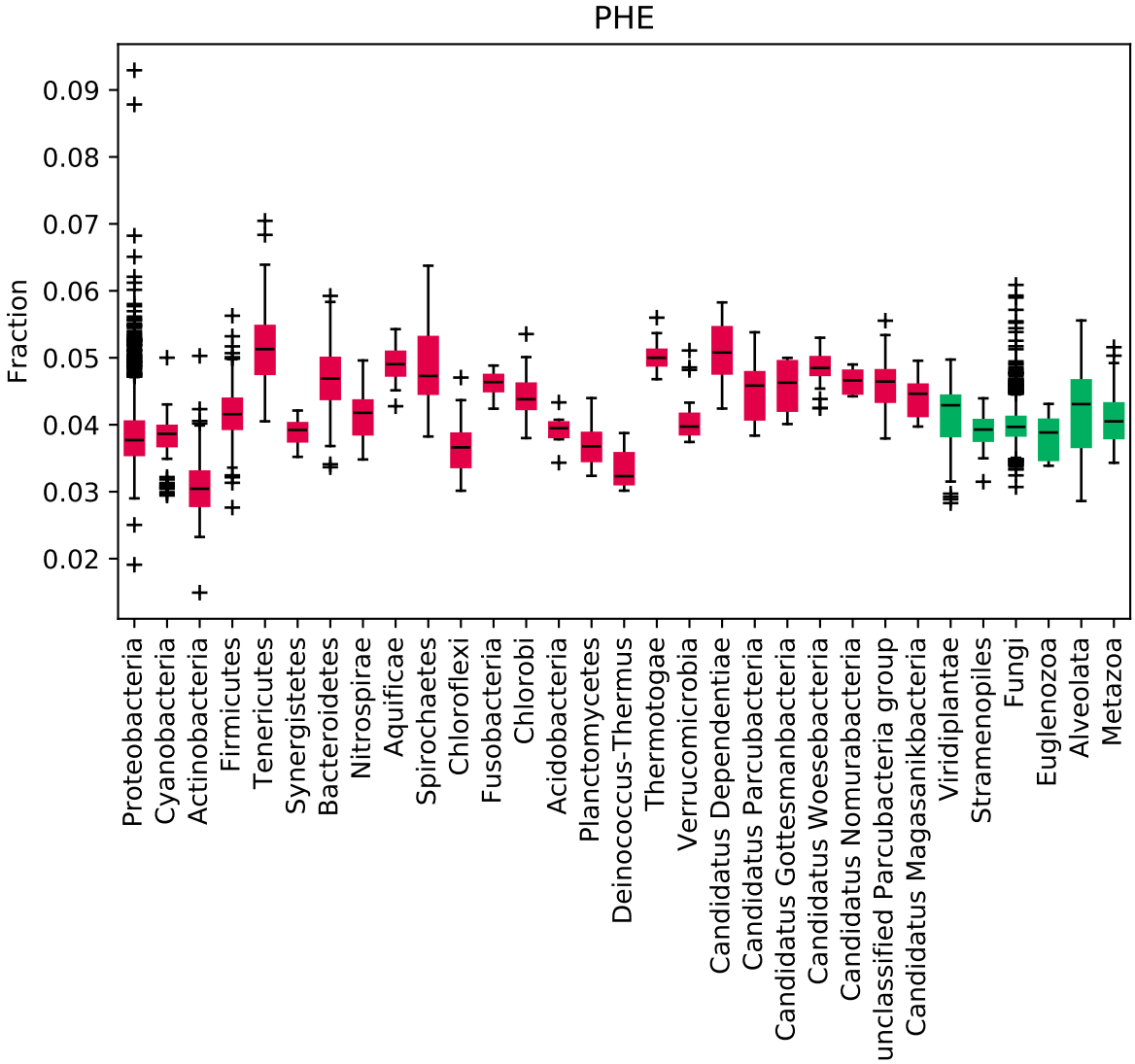


Figure S7. Frequency of phenylalanine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

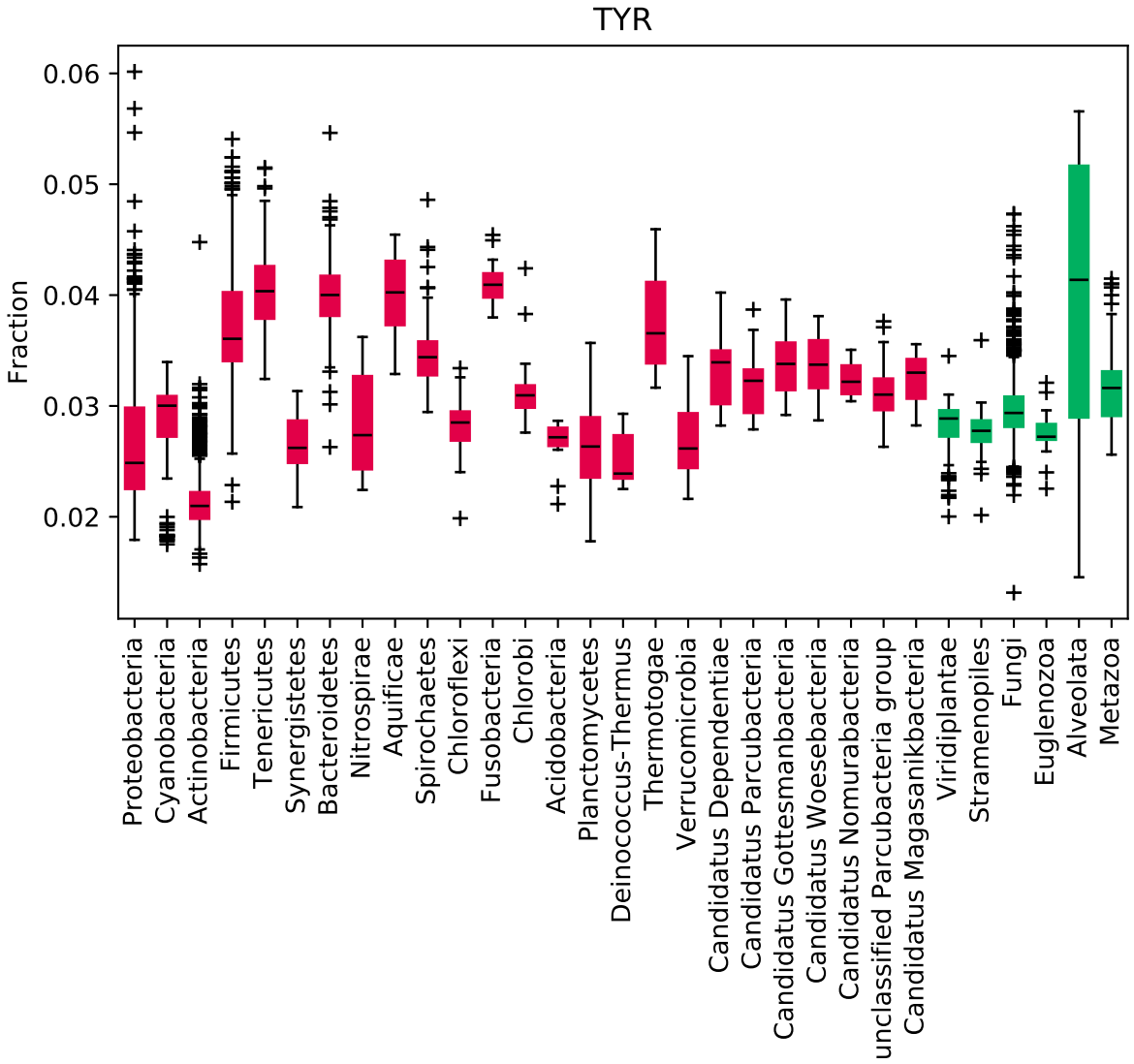


Figure S8. Frequency of tyrosine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

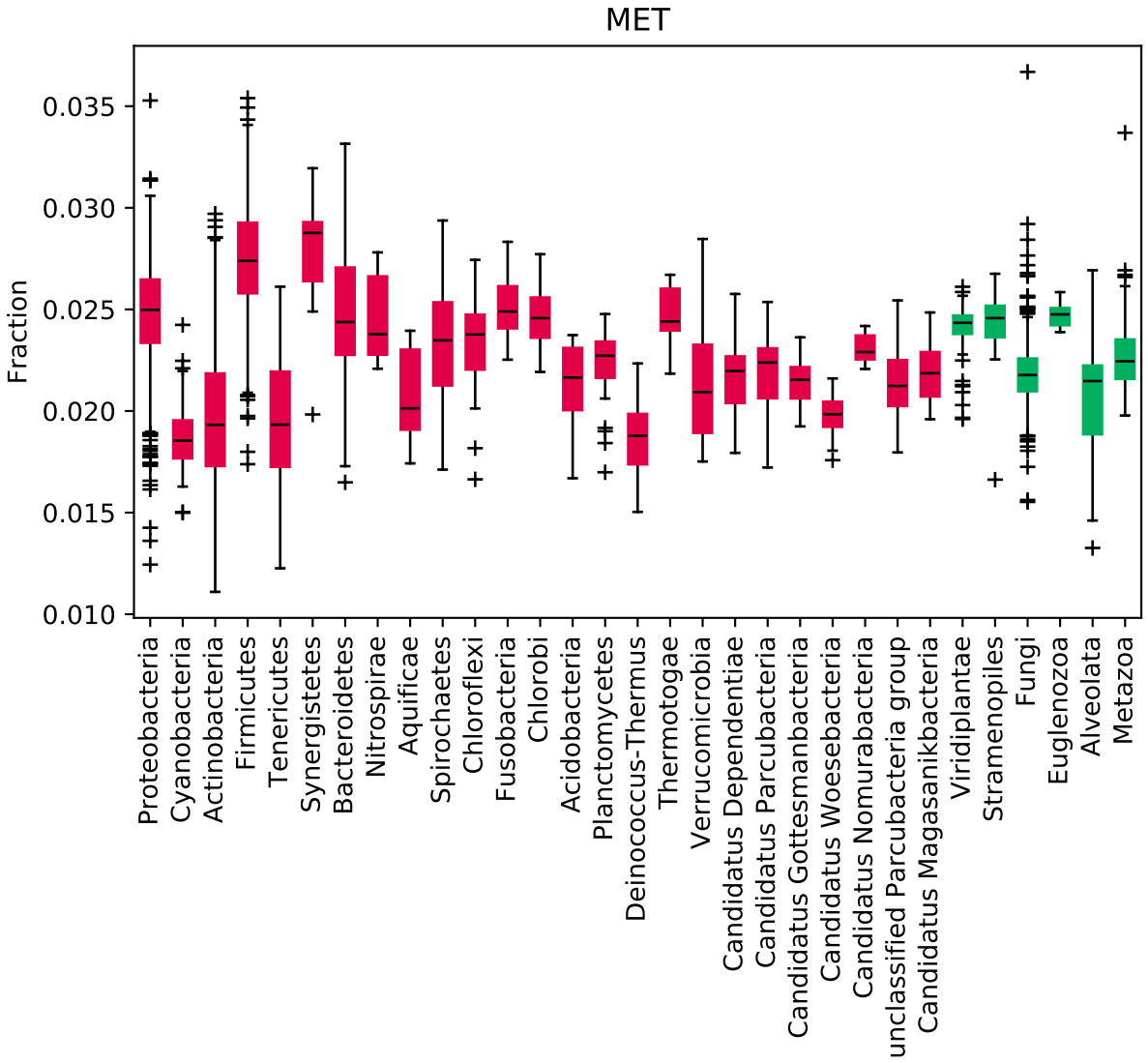


Figure S9. Frequency of methionine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

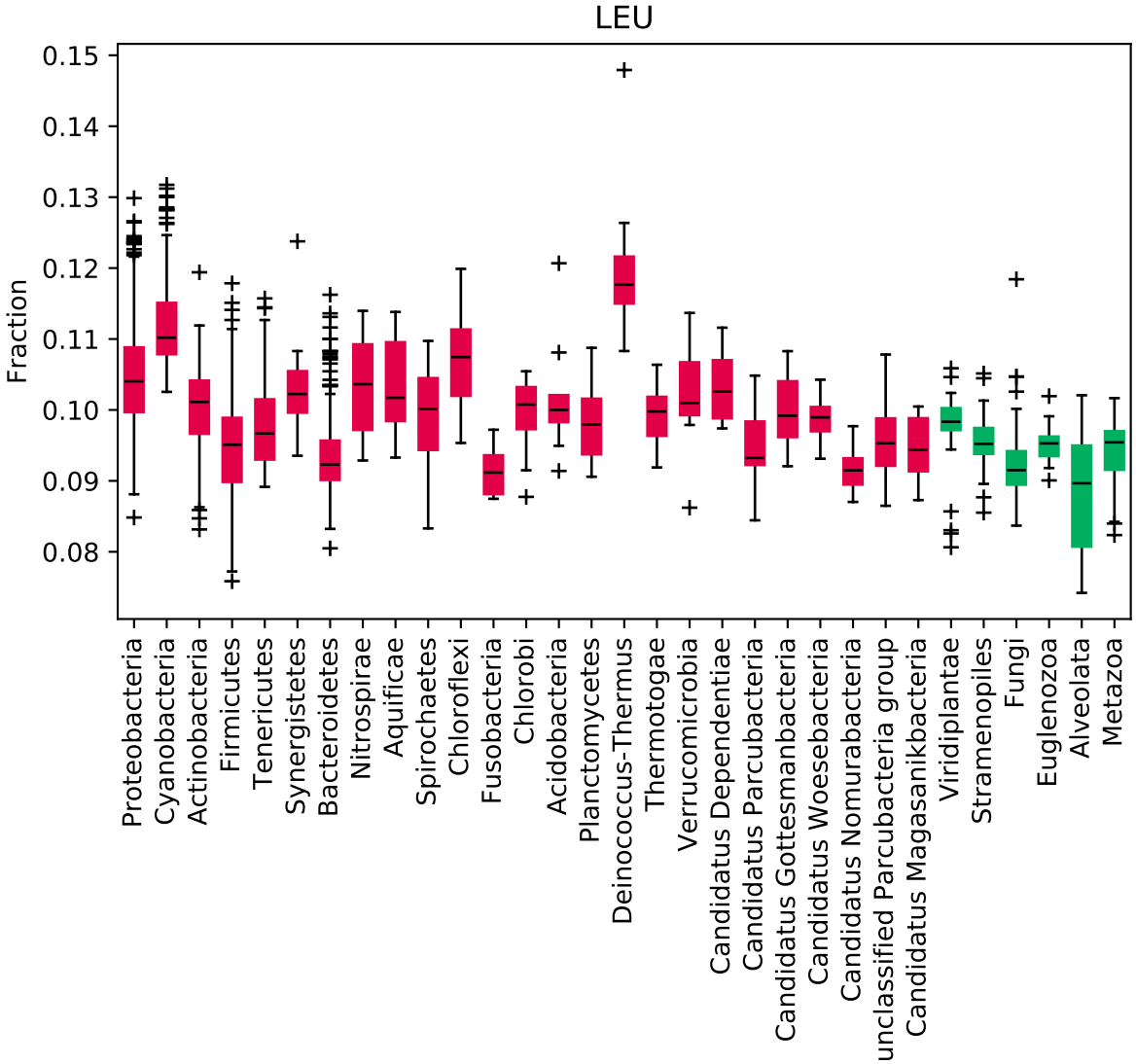


Figure S10. Frequency of leucine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

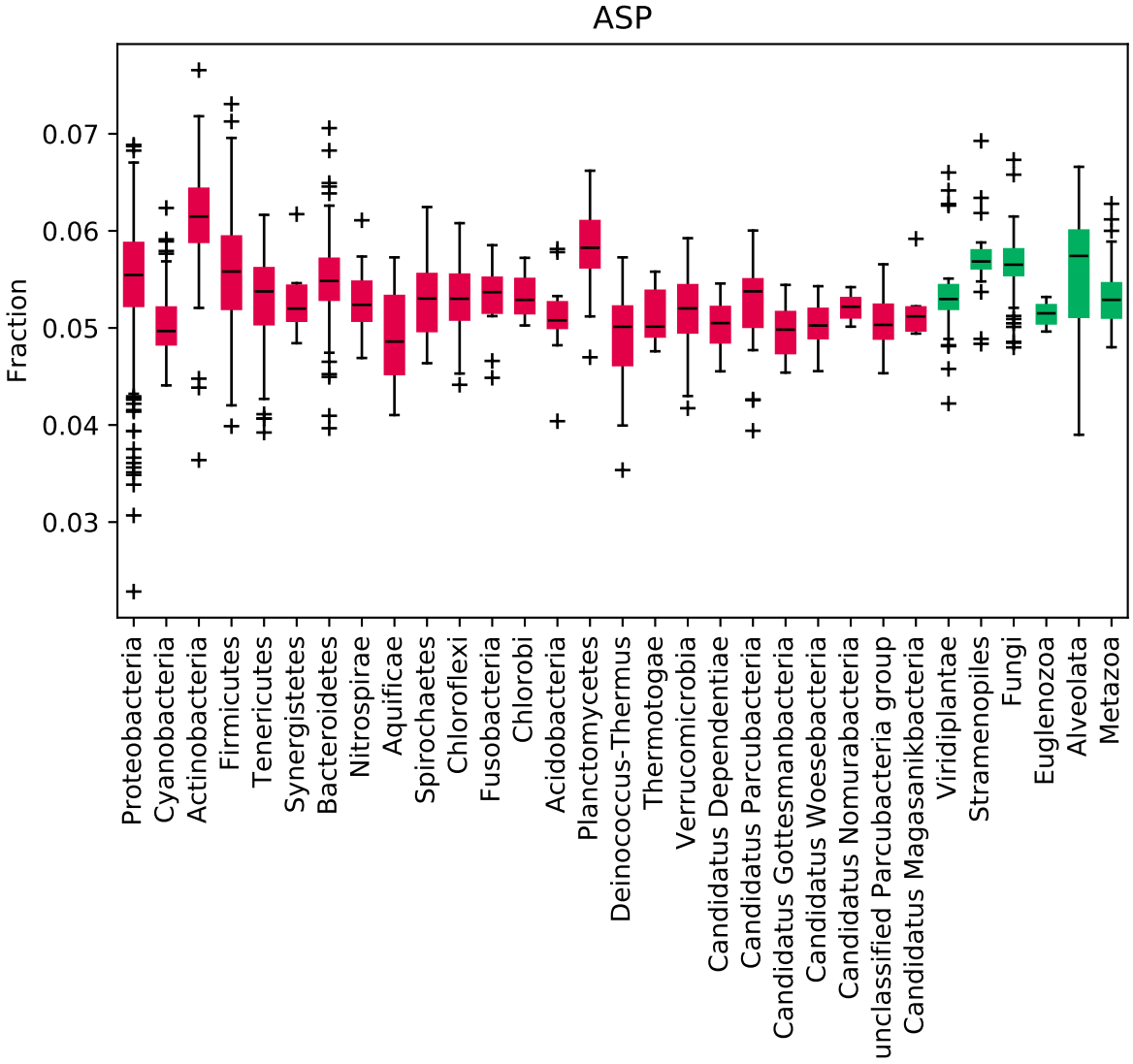


Figure S11. Frequency of aspartic acid in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

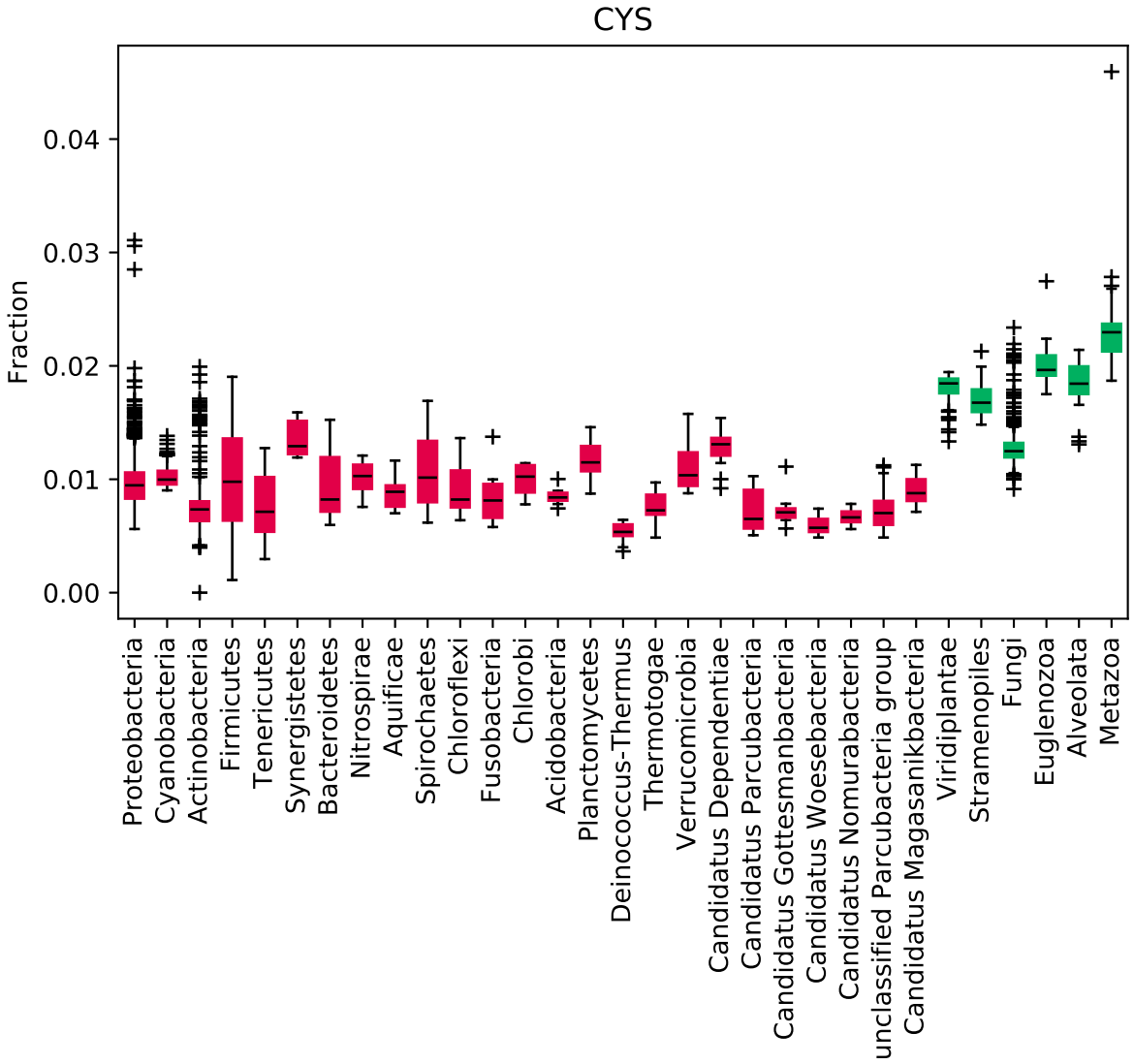


Figure S12. Frequency of cysteine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

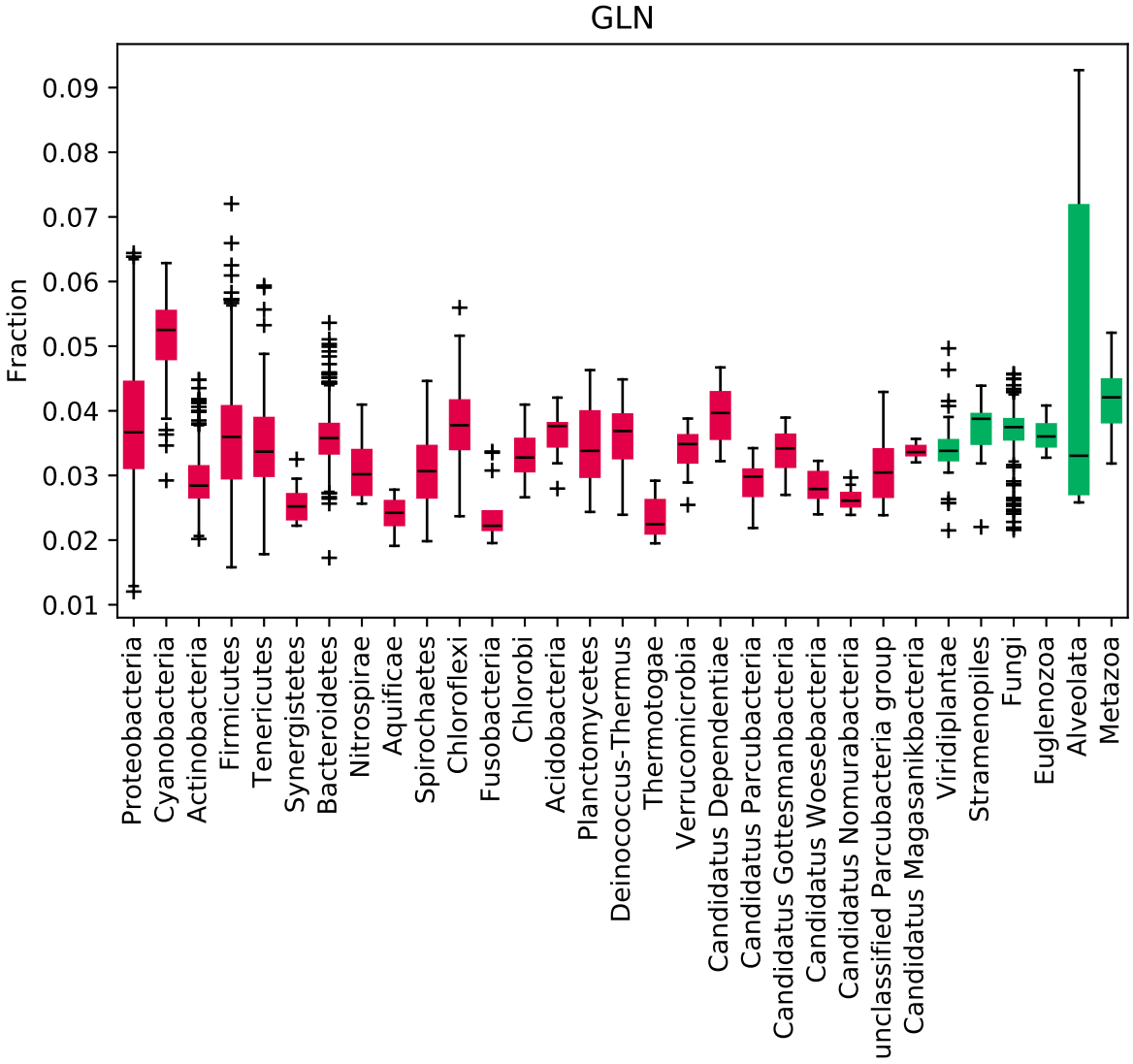


Figure S13. Frequency of glutamine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

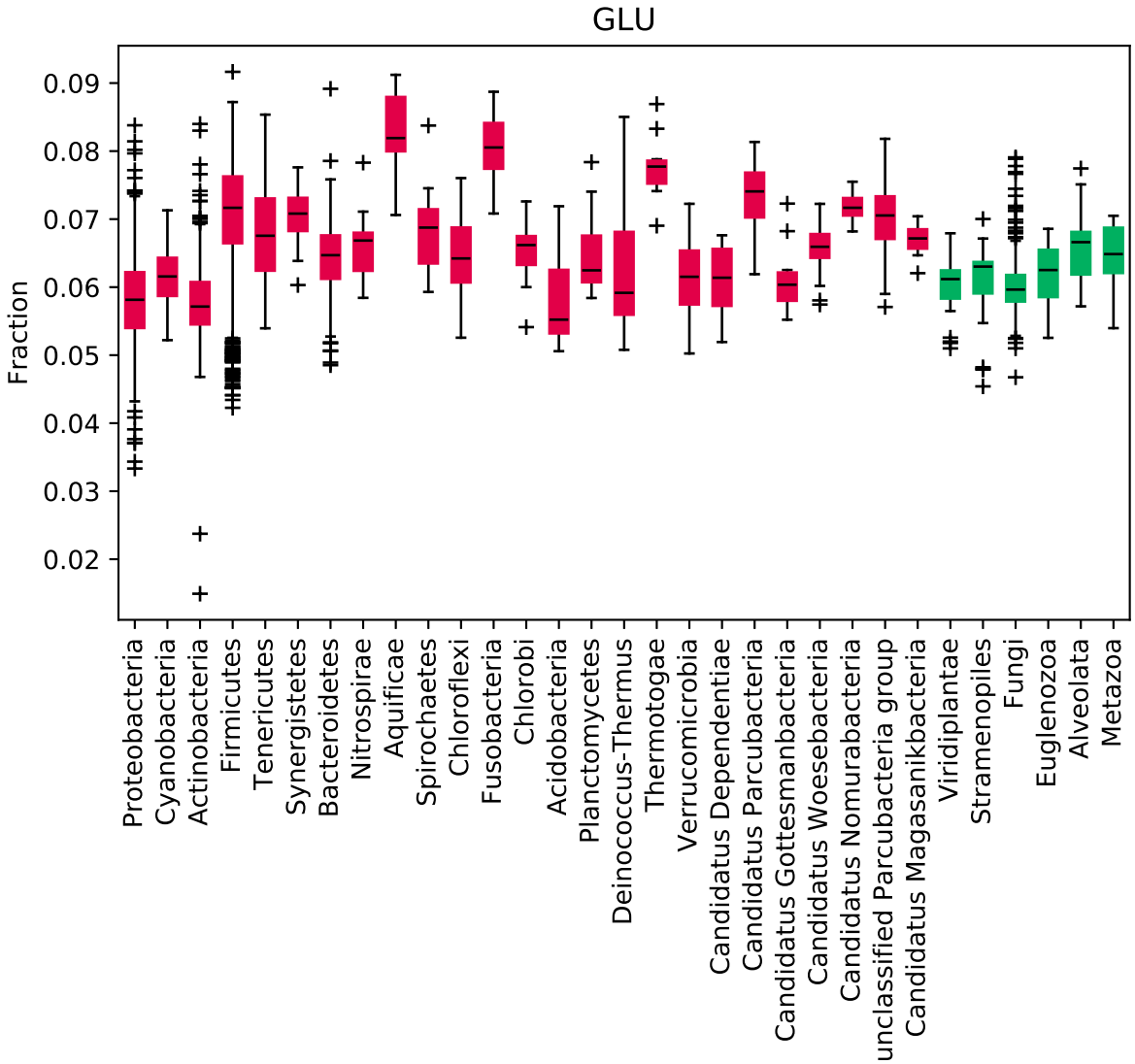


Figure S14. Frequency of glutamic acid in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

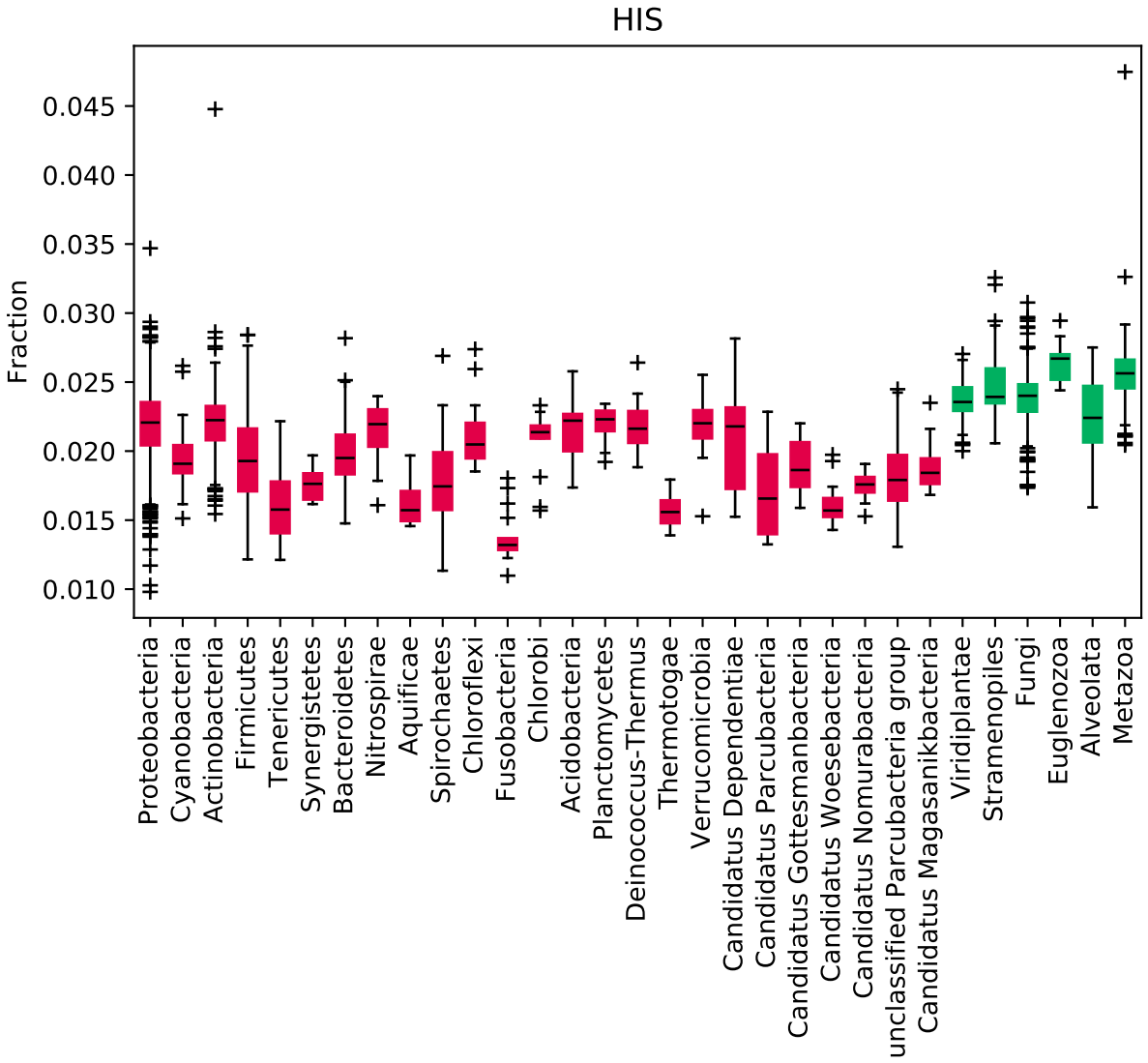


Figure S15. Frequency of histidine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

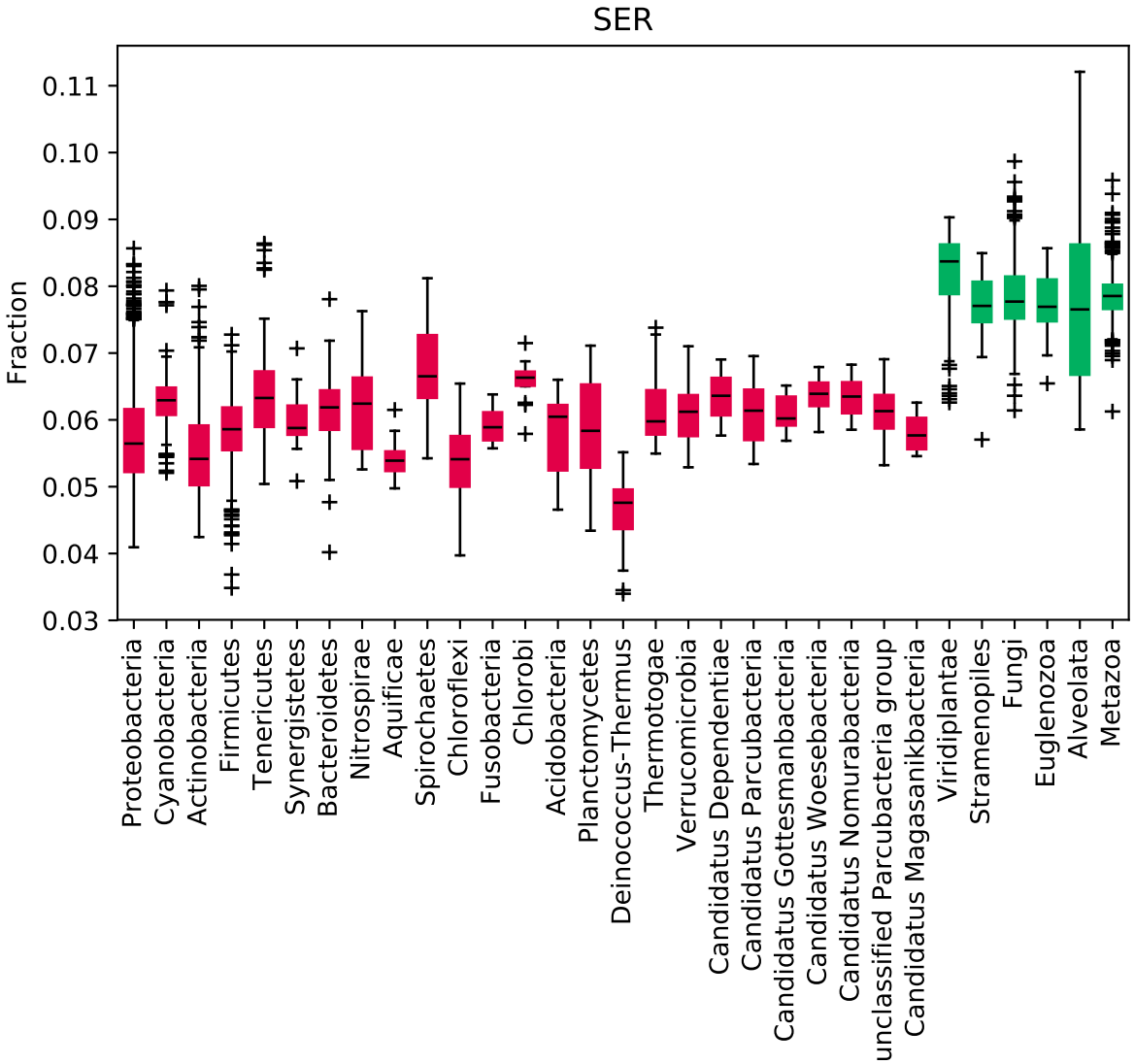


Figure S16. Frequency of serine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

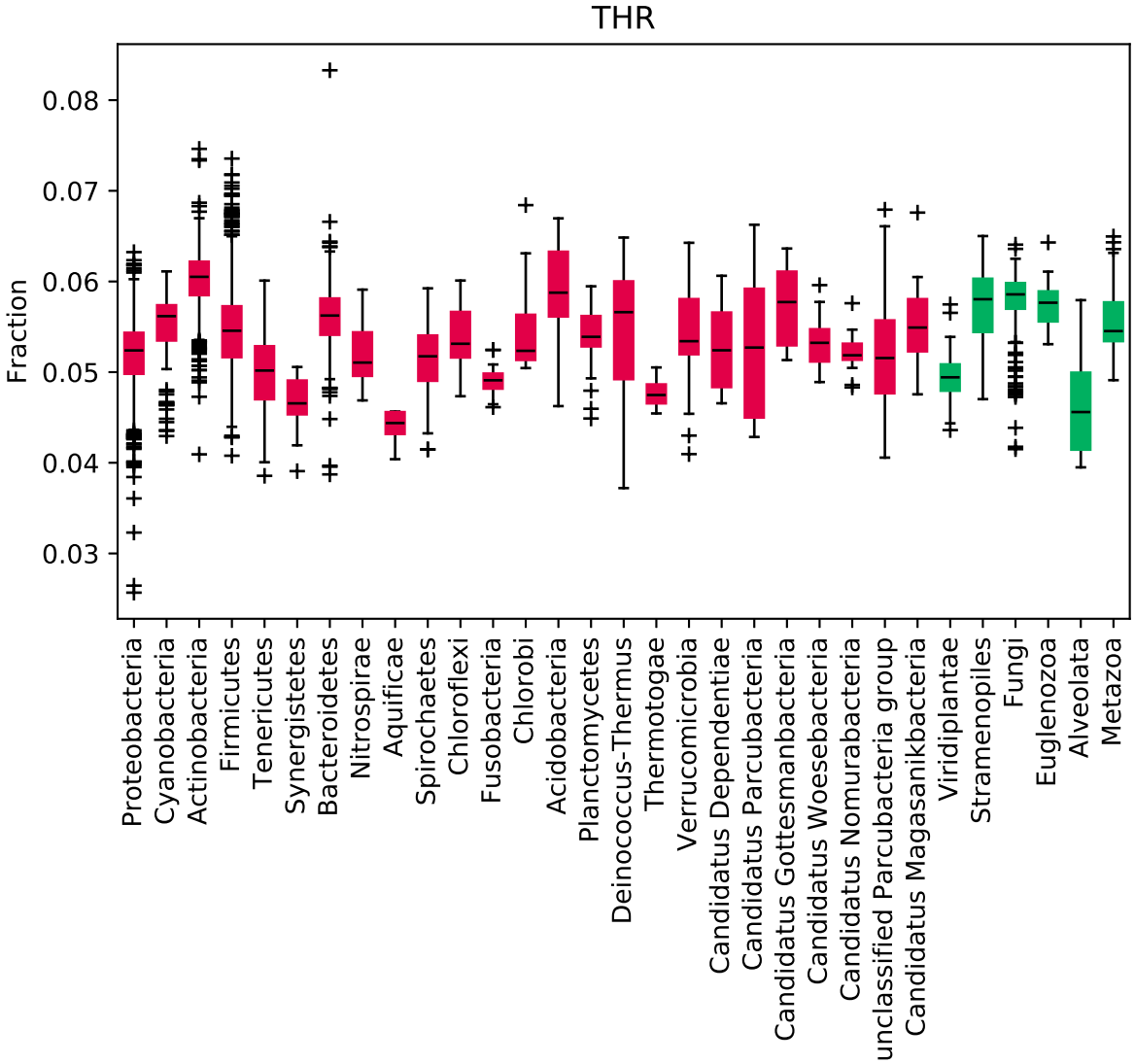


Figure S17. Frequency of threonine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

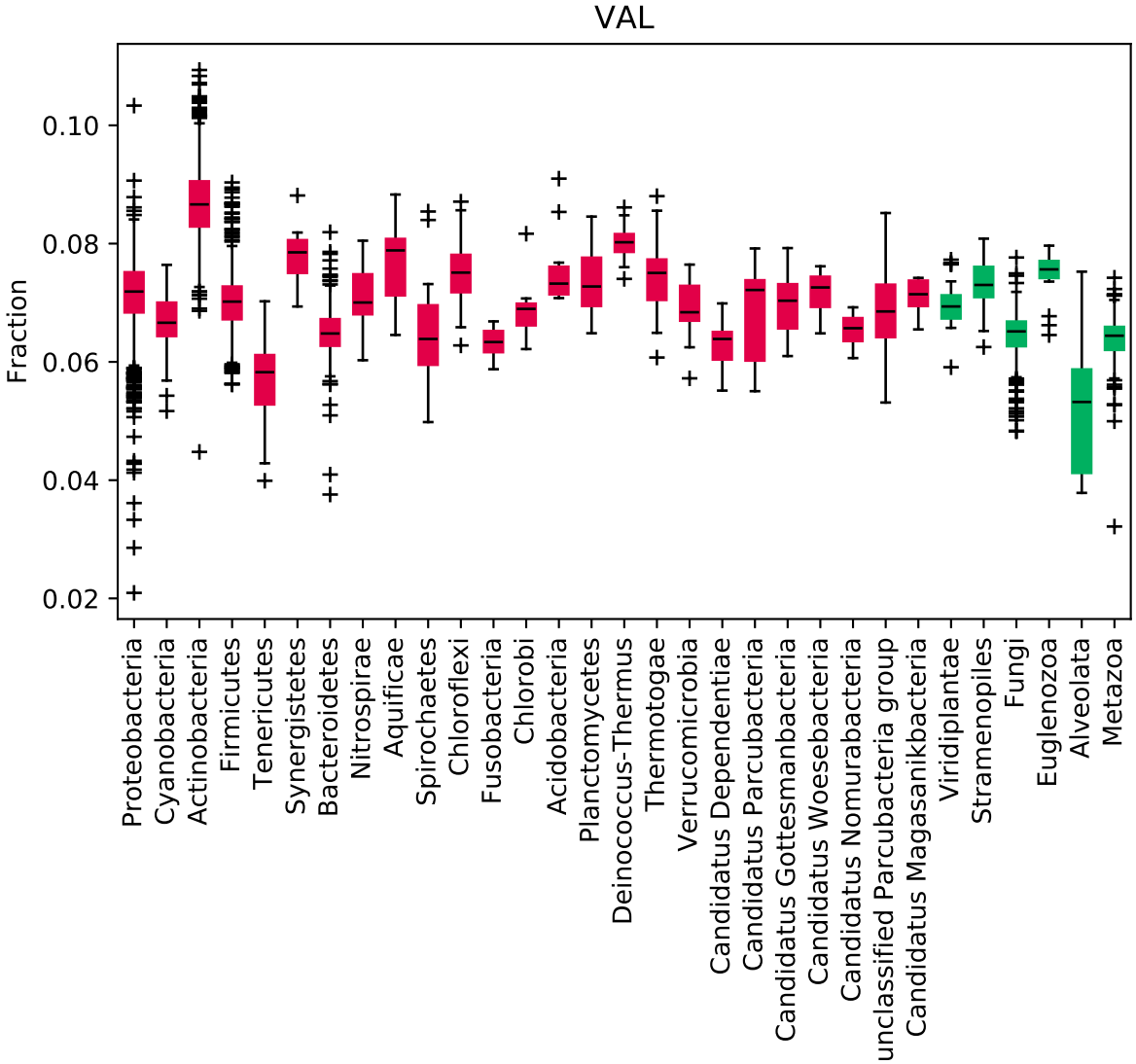


Figure S18. Frequency of valine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

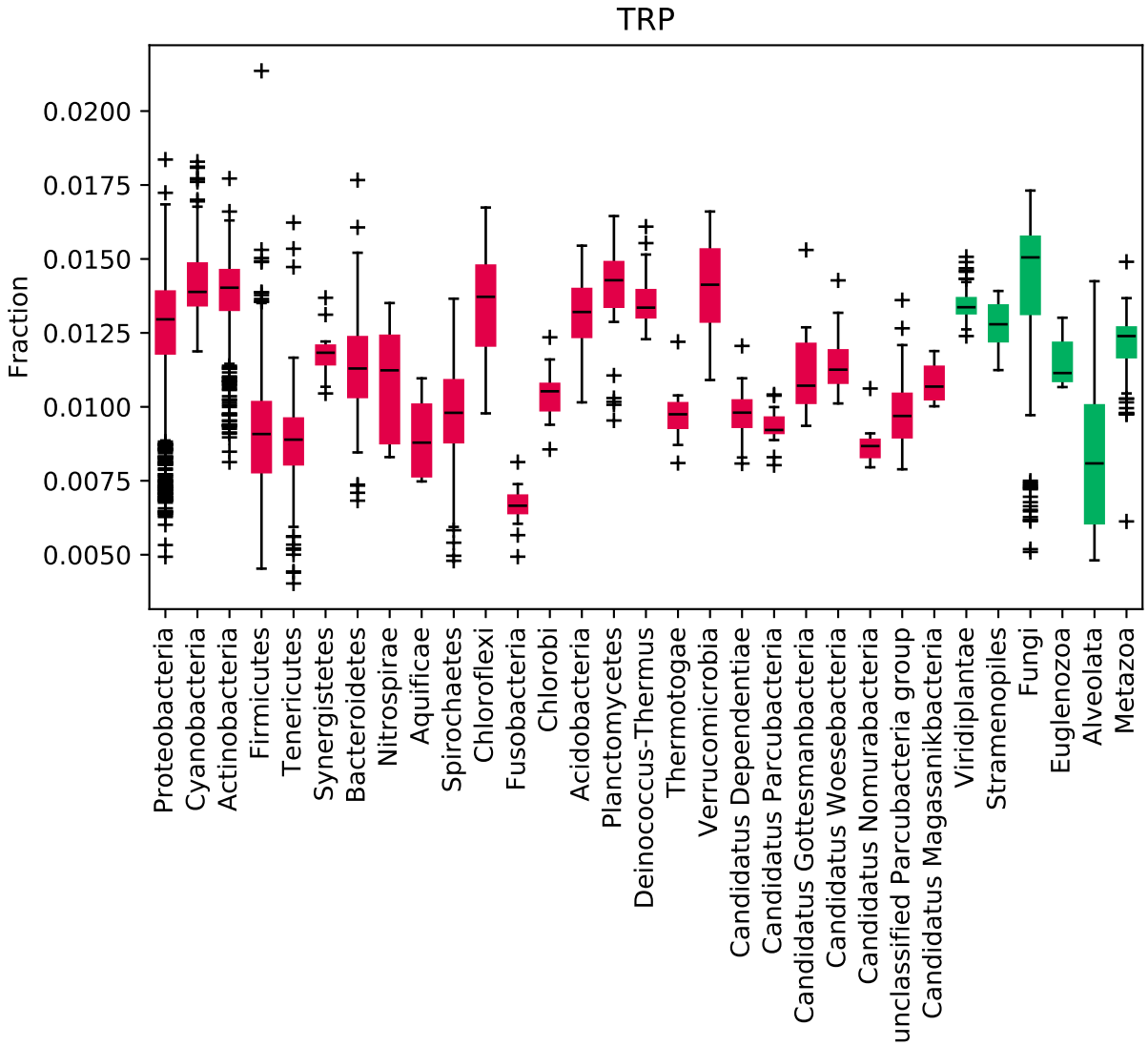


Figure S19. Frequency of tryptophan in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

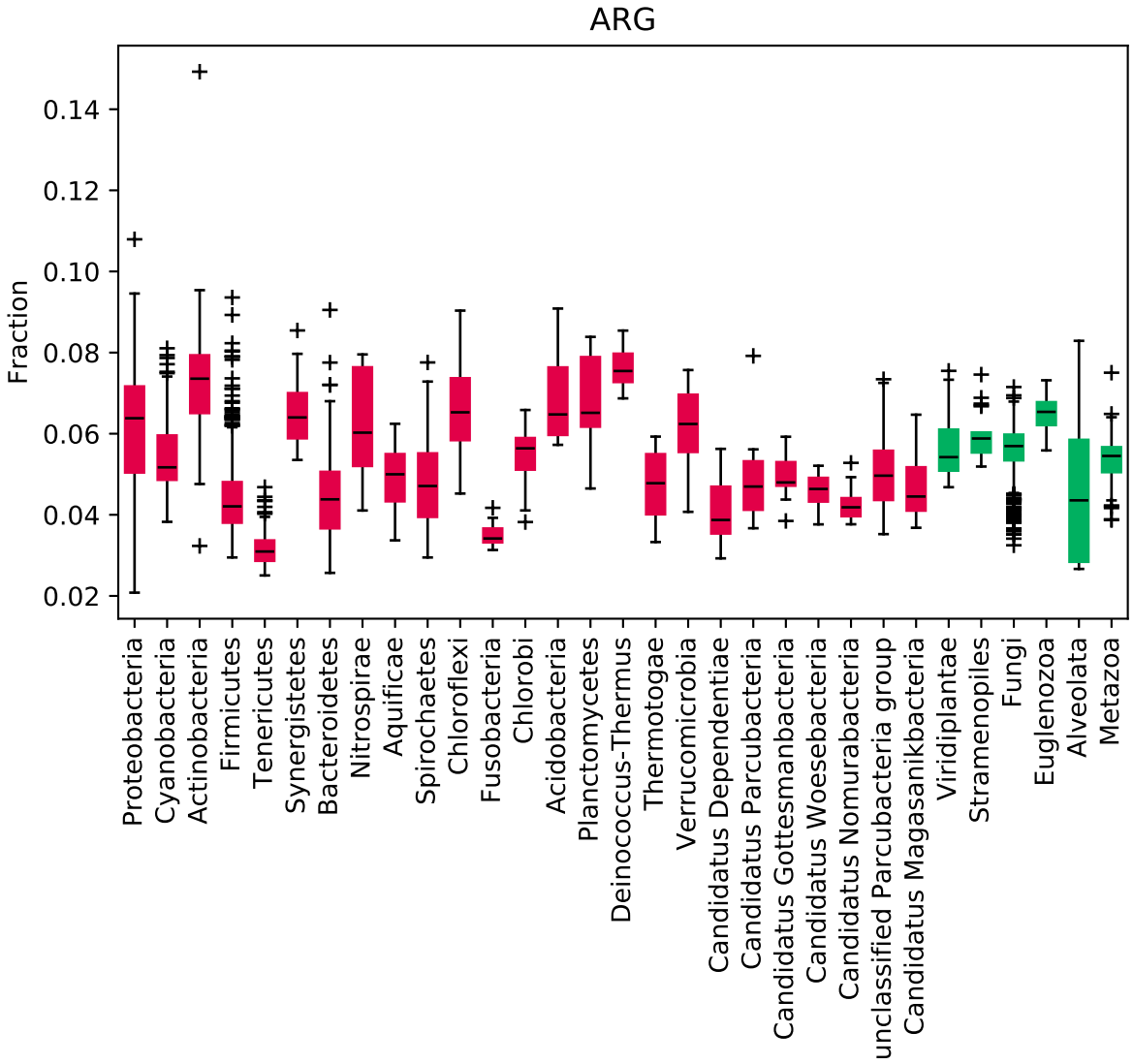


Figure S20. Frequency of arginine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

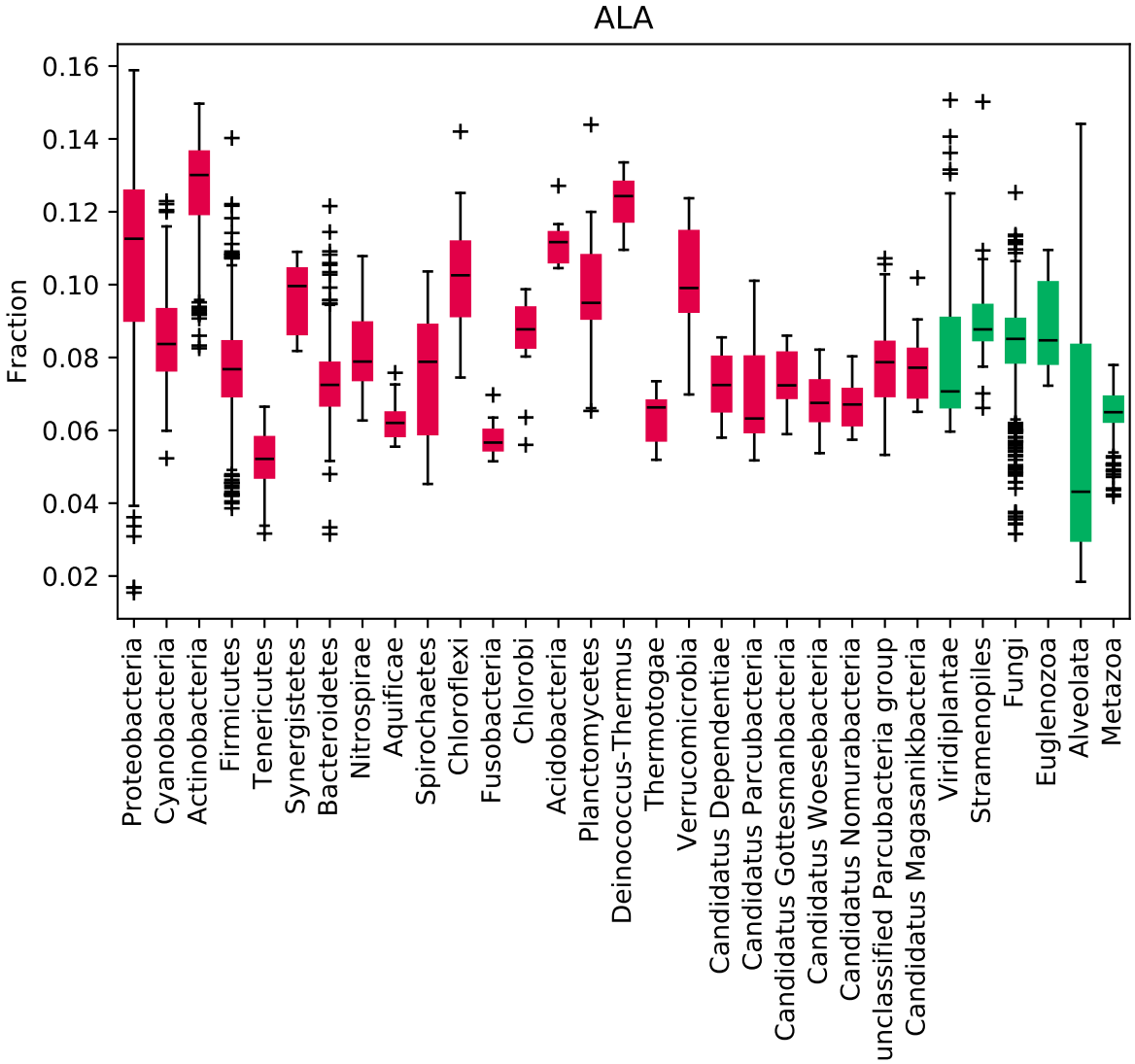


Figure S21. Frequency of alanine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

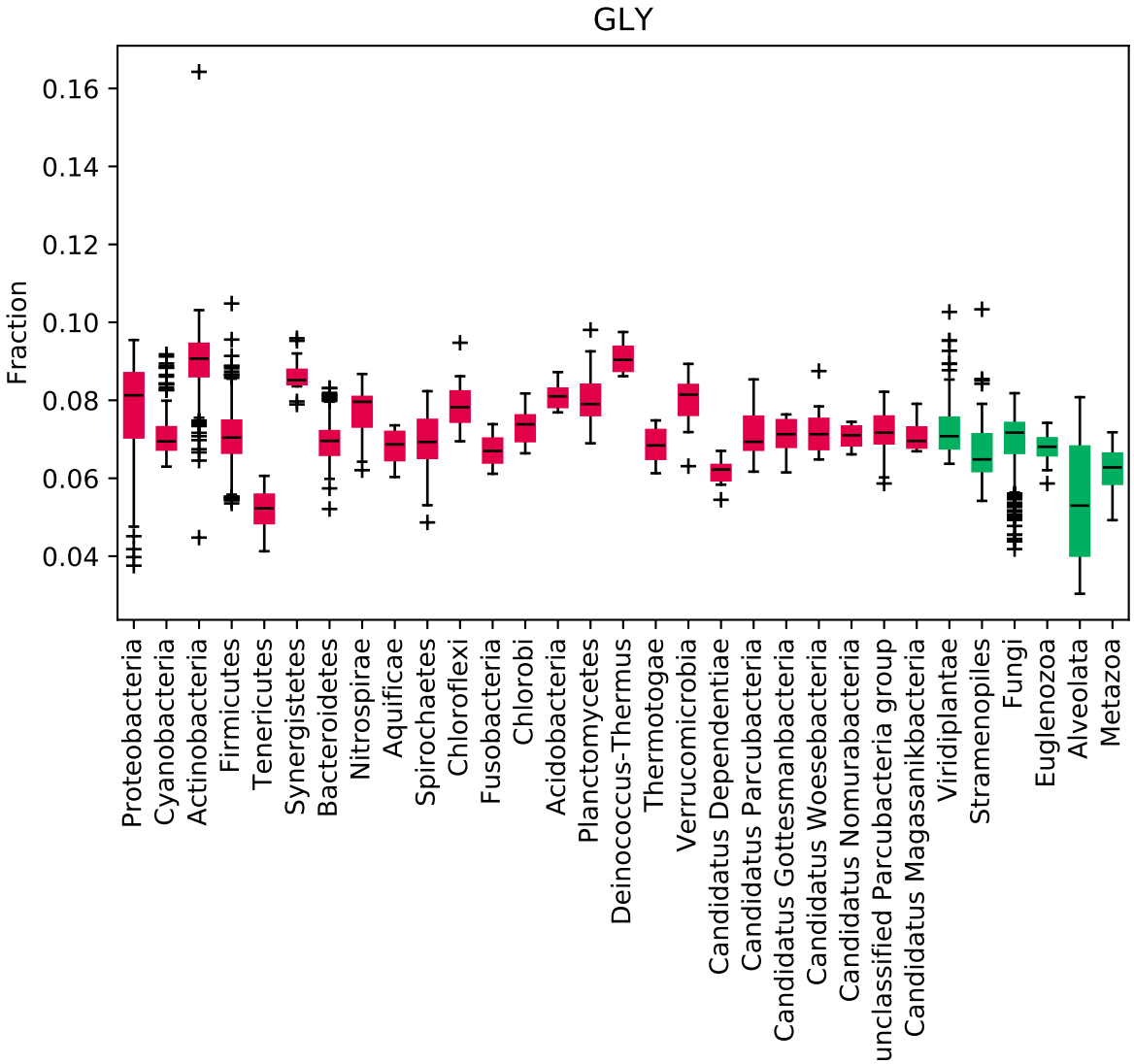


Figure S22. Frequency of glycine in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.

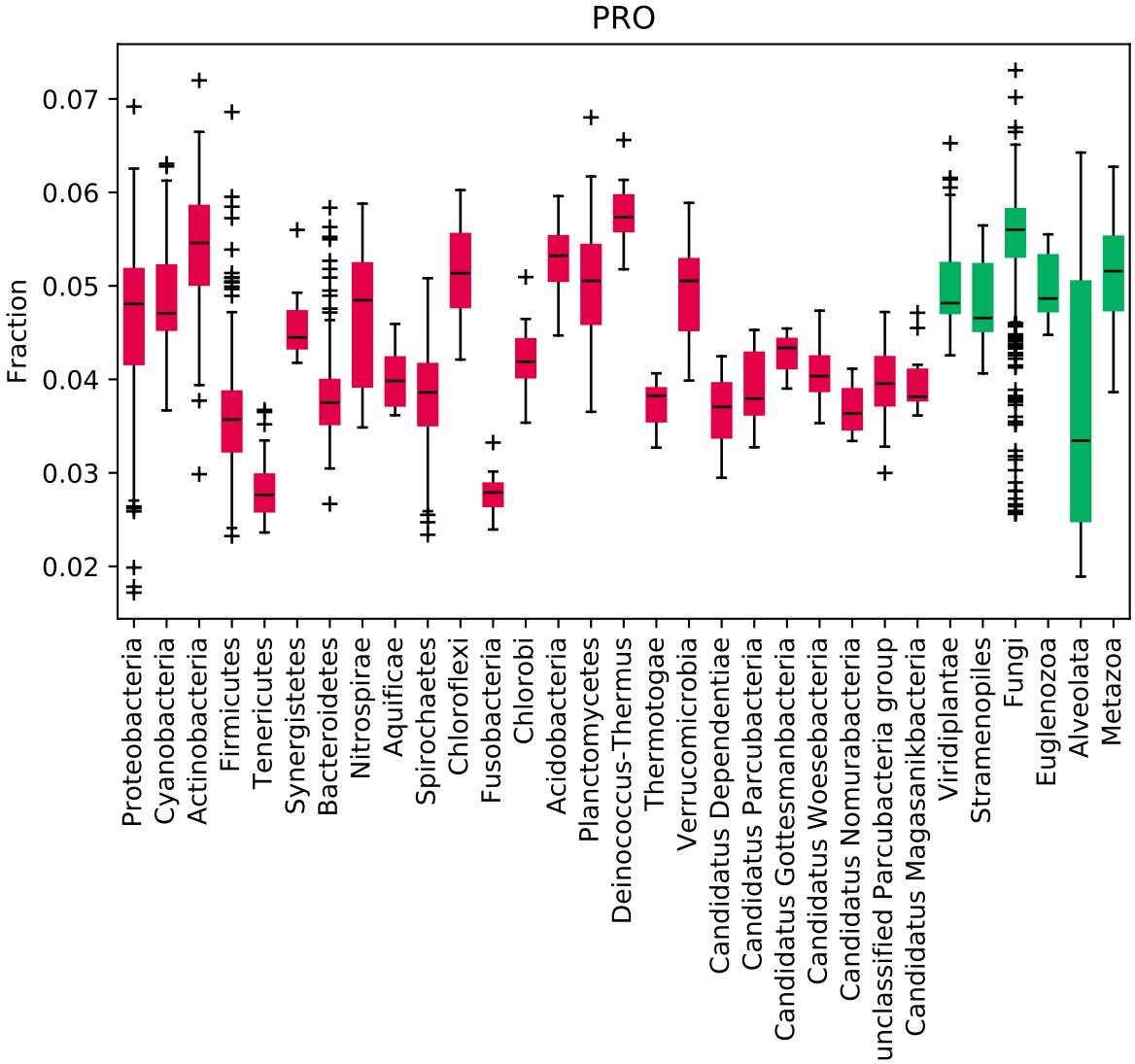
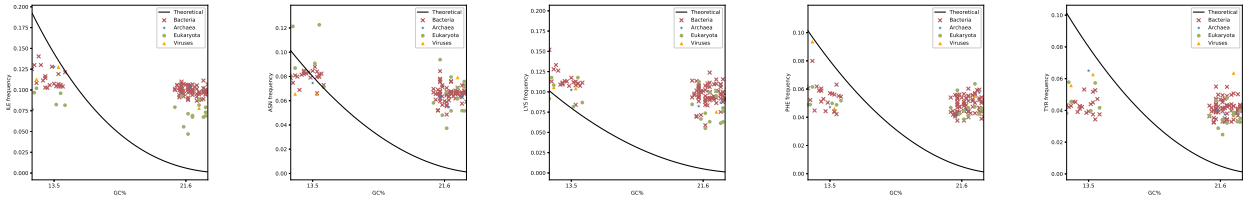
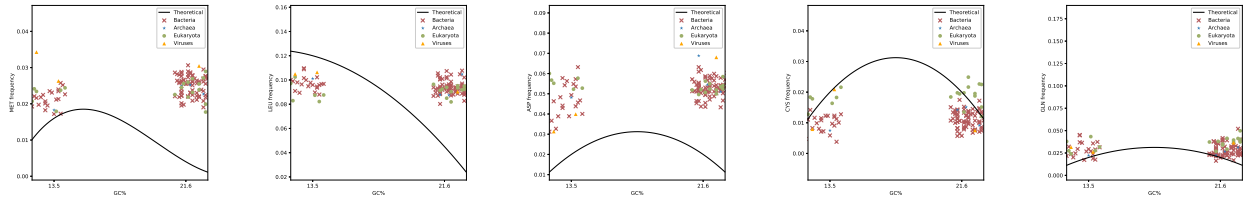


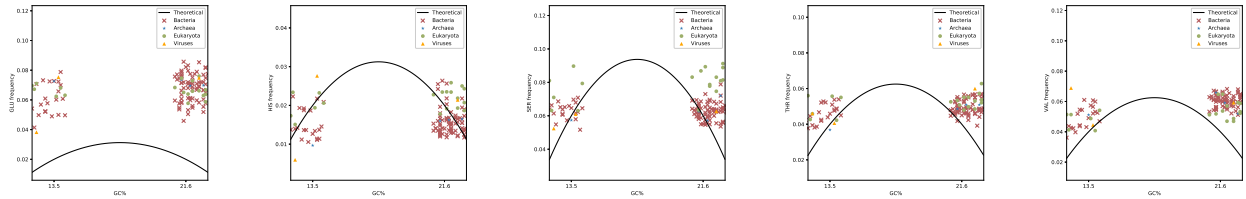
Figure S23. Frequency of proline in complete genomes grouped by phylum. Bacterial groups are red, eukaryotic green and archaeal blue.



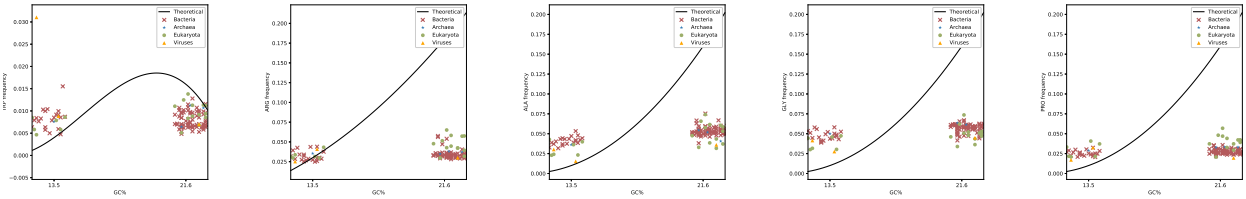
(a) 11% Ile **(b)** 17% Asn **(c)** 17% Lys **(d)** 17% Phe **(e)** 17% Tyr



(f) 33% Met **(g)** 39% Leu **(h)** 50% Asp **(i)** 50% Cys **(j)** 50% Gln



(k) 50% Glu **(l)** 50% His **(m)** 50% Ser **(n)** 50% Thr **(o)** 50% Val



(p) 67% Trp **(q)** 72% Arg **(r)** 83% Ala **(s)** 83% Gly **(t)** 83% Pro

Figure S24. Frequency of different vs GC of the genomes amino acids are sorted by the GC content of the codons. The number represent the fraction of GC among the codons.