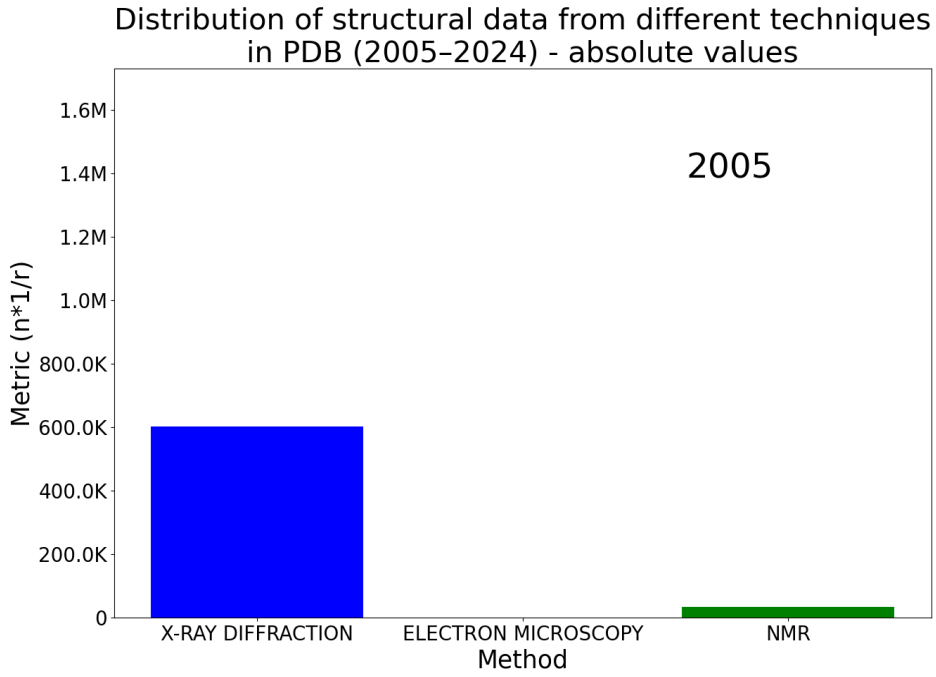
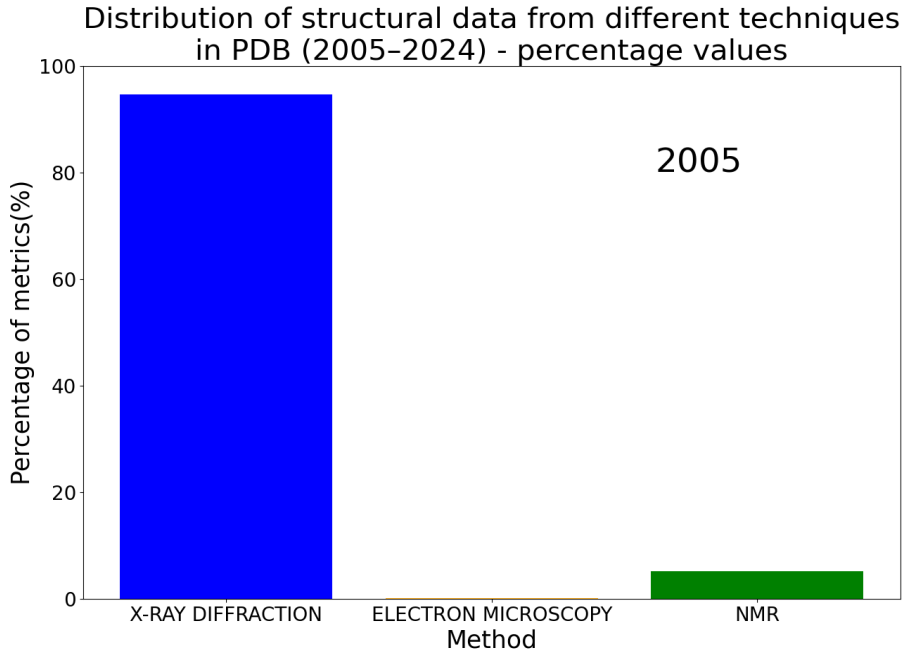
Justyna Kowalska

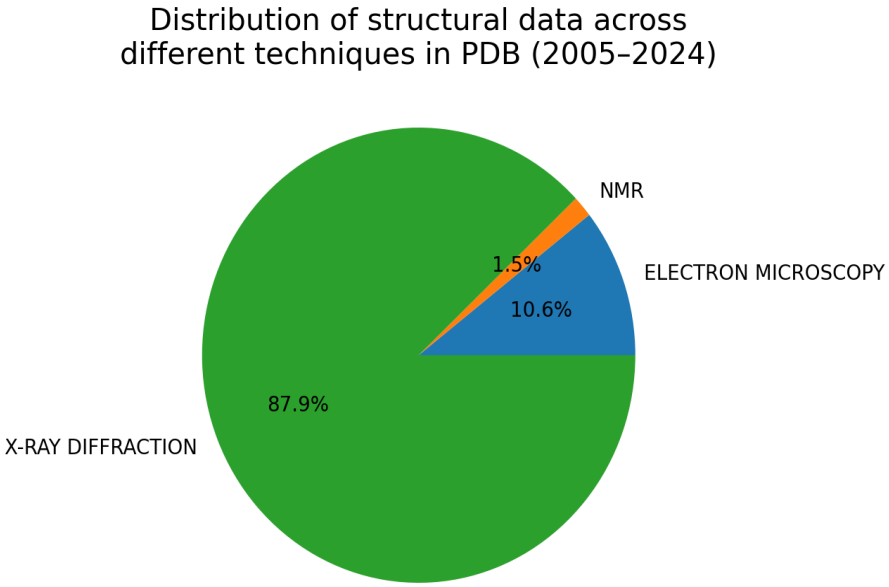
REPORT

Architecture of large projects in bioinformatics

Exercise 1





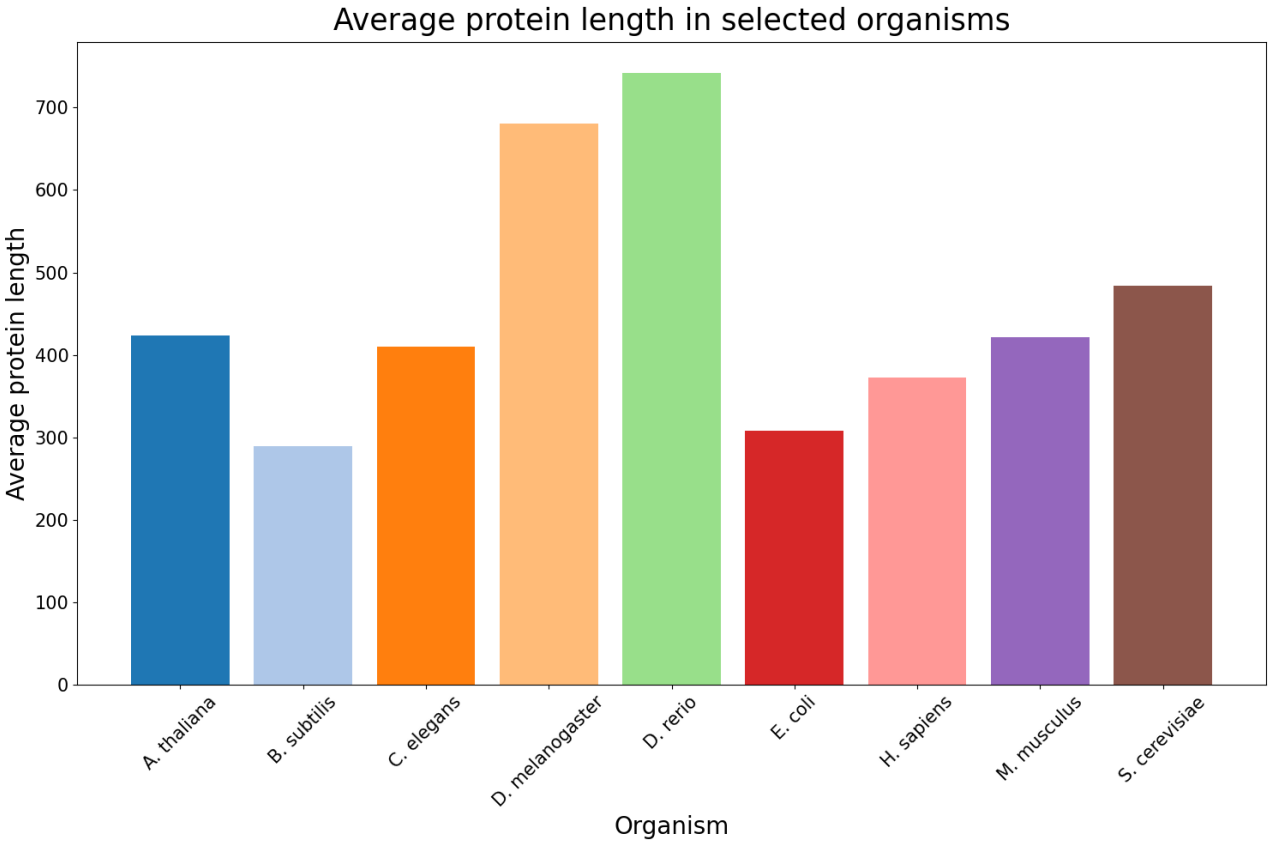


Exercise 2

Table 2: Protein length by organism

Index	Average length	STD
A. thaliana	423.31	15.05
B. subtilis	289.78	12.87
C. elegans	410.05	12.04
D. melanogaster	680.60	96.69
D. rerio	742.30	23.99
E. coli	307.66	6.59
H. sapiens	372.72	21.69
M. musculus	421.80	42.71
S. cerevisiae	484.33	12.05

Values are in [aa]


Average protein length in selected organisms
Average protein length
0
100
200
300
400
500
600
700
A. thaliana
B. subtilis
C. elegans
D. melanogaster
D. rerio
E. coli
H. sapiens
M. musculus
S. cerevisiae
Organism

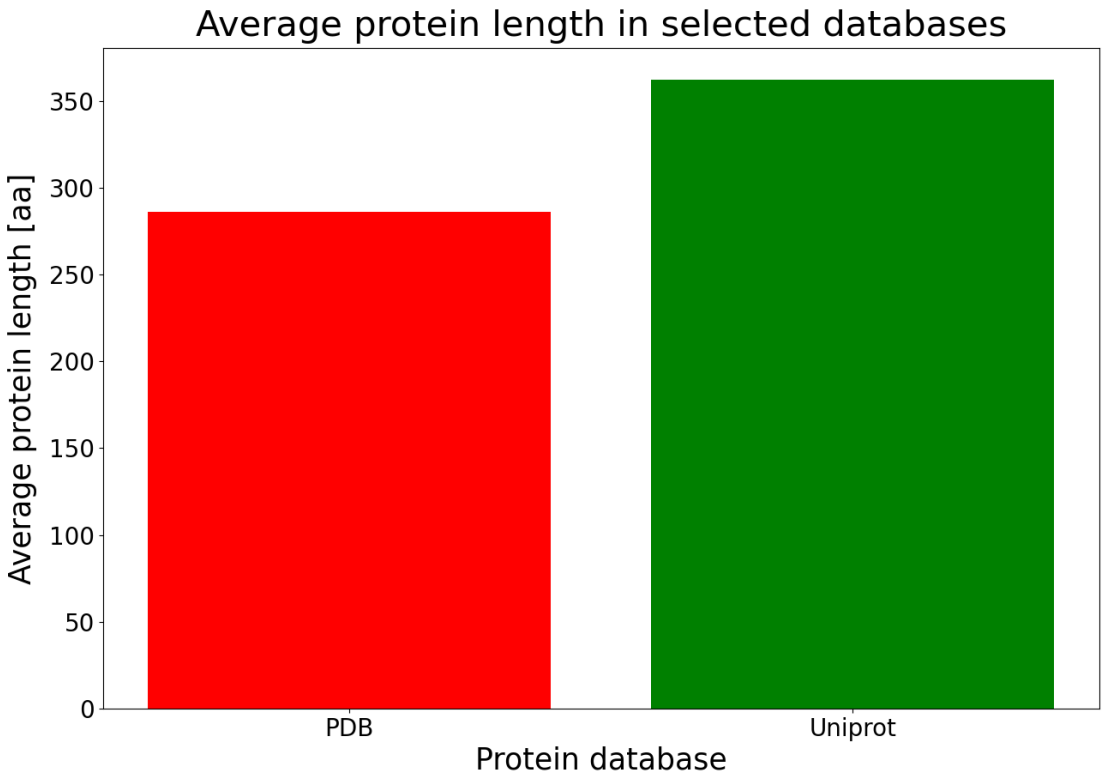

Average protein length in selected databases
Average protein length [aa]
0
50
100
150
200
250
300
350
PDB
Uniprot
Protein database

Table 4: Protein length by taxonomy

Index	Average length	STD
Archaea	278.25	6.95
Bacteria	322.92	7.81
Eukaryota	473.45	16.43
Viruses	572.71	127.92

Values are in [aa]

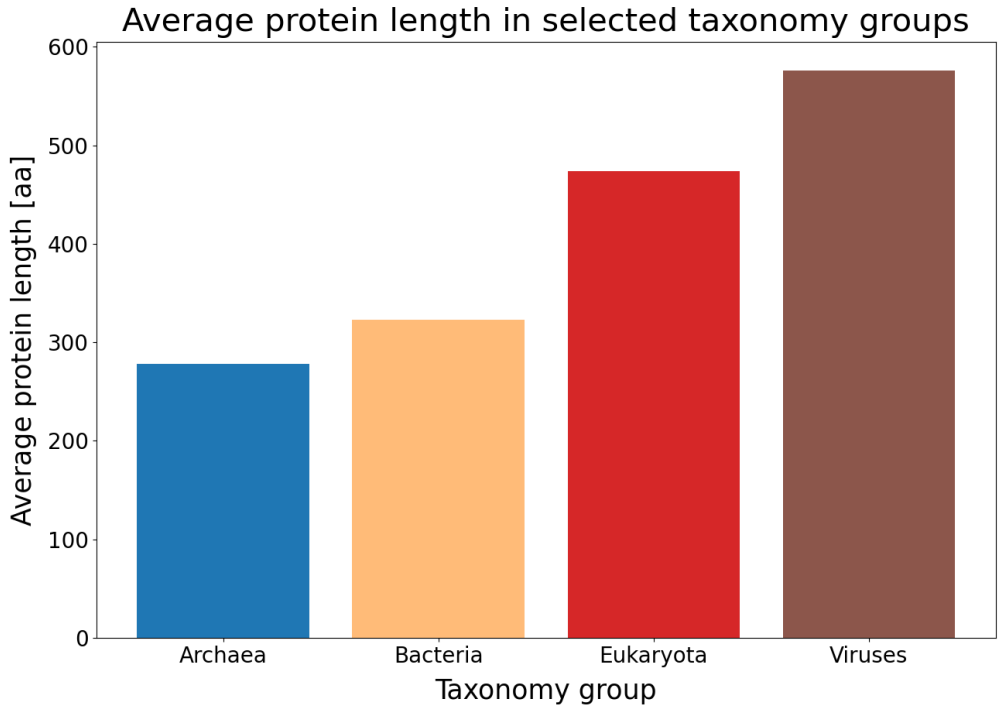


Table 1: Average amino acid content by organism

Index	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A. thaliana	6.3 ± 0.061	1.9 ± 0.029	5.4 ± 0.043	6.8 ± 0.085	4.3 ± 0.044	6.3 ± 0.064	2.3 ± 0.027	5.3 ± 0.043	6.4 ± 0.06	9.6 ± 0.097	2.4 ± 0.024	4.4 ± 0.042	4.8 ± 0.061	3.5 ± 0.05	5.4 ± 0.048	9.2 ± 0.076	5.1 ± 0.043	6.6 ± 0.043	1.2 ± 0.025	2.8 ± 0.042
B. subtilis	7.7 ± 0.066	0.79 ± 0.022	5.2 ± 0.049	7.3 ± 0.083	4.5 ± 0.053	6.9 ± 0.069	2.3 ± 0.041	7.4 ± 0.066	7.1 ± 0.079	9.7 ± 0.076	2.8 ± 0.034	4.0 ± 0.052	3.7 ± 0.05	3.8 ± 0.055	4.1 ± 0.064	6.3 ± 0.075	5.4 ± 0.058	6.7 ± 0.059	1.0 ± 0.025	3.5 ± 0.048
C. elegans	6.3 ± 0.084	2.1 ± 0.062	5.3 ± 0.059	6.5 ± 0.11	4.8 ± 0.065	5.4 ± 0.083	2.3 ± 0.051	6.2 ± 0.058	6.3 ± 0.095	8.6 ± 0.087	2.7 ± 0.043	4.9 ± 0.057	4.9 ± 0.079	4.1 ± 0.067	5.1 ± 0.079	8.1 ± 0.1	5.9 ± 0.069	6.2 ± 0.049	1.1 ± 0.029	3.2 ± 0.051
D. melanogaster	7.4 ± 0.1	1.9 ± 0.35	5.2 ± 0.07	6.6 ± 0.2	3.3 ± 0.054	6.2 ± 0.11	2.6 ± 0.034	4.8 ± 0.066	5.5 ± 0.17	8.7 ± 0.21	2.2 ± 0.059	4.7 ± 0.078	5.8 ± 0.25	5.4 ± 0.088	5.5 ± 0.067	8.6 ± 0.1	5.9 ± 0.13	5.9 ± 0.093	0.92 ± 0.039	2.8 ± 0.035
D. rerio	6.3 ± 0.052	2.1 ± 0.07	5.3 ± 0.043	7.2 ± 0.098	3.4 ± 0.052	6.0 ± 0.084	2.7 ± 0.028	4.5 ± 0.046	6.0 ± 0.085	9.2 ± 0.097	2.3 ± 0.032	4.0 ± 0.043	5.8 ± 0.089	5.0 ± 0.082	5.5 ± 0.055	9.2 ± 0.14	5.8 ± 0.15	6.1 ± 0.047	1.0 ± 0.023	2.6 ± 0.043
E. coli	9.5 ± 0.086	1.2 ± 0.022	5.1 ± 0.055	5.8 ± 0.086	3.9 ± 0.052	7.4 ± 0.063	2.3 ± 0.038	6.0 ± 0.054	4.4 ± 0.061	1.1e+01 ± 0.084	2.8 ± 0.031	3.9 ± 0.044	4.4 ± 0.05	4.4 ± 0.061	5.5 ± 0.066	5.8 ± 0.053	5.4 ± 0.045	7.1 ± 0.057	1.5 ± 0.038	2.8 ± 0.044
H. sapiens	6.9 ± 0.095	2.2 ± 0.098	4.8 ± 0.049	7.2 ± 0.11	3.6 ± 0.055	6.5 ± 0.086	2.6 ± 0.028	4.3 ± 0.077	5.8 ± 0.11	9.9 ± 0.084	2.2 ± 0.034	3.6 ± 0.064	6.2 ± 0.12	4.8 ± 0.072	5.6 ± 0.063	8.4 ± 0.086	5.5 ± 0.047	6.0 ± 0.056	1.2 ± 0.027	2.6 ± 0.046
M. musculus	6.8 ± 0.065	2.3 ± 0.066	4.8 ± 0.047	7.0 ± 0.11	3.7 ± 0.068	6.4 ± 0.077	2.6 ± 0.049	4.4 ± 0.082	5.7 ± 0.12	1e+01 ± 0.15	2.2 ± 0.044	3.6 ± 0.055	6.2 ± 0.12	4.8 ± 0.099	5.6 ± 0.058	8.6 ± 0.1	5.4 ± 0.081	6.1 ± 0.12	1.2 ± 0.023	2.7 ± 0.041
S. cerevisiae	5.5 ± 0.066	1.3 ± 0.02	5.8 ± 0.049	6.5 ± 0.073	4.4 ± 0.044	5.0 ± 0.075	2.2 ± 0.029	6.6 ± 0.05	7.3 ± 0.077	9.5 ± 0.076	2.1 ± 0.024	6.2 ± 0.063	4.4 ± 0.049	3.9 ± 0.06	4.4 ± 0.044	9.0 ± 0.13	5.9 ± 0.1	5.6 ± 0.056	1.0 ± 0.02	3.4 ± 0.035

Values are in 10^{-2} format, ± indicates standard deviation via bootstrap 100

Table 3: Average amino acid content by taxonomy

Index	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Archaea	8.4 ± 0.096	0.89 ± 0.03	6.5 ± 0.075	8.1 ± 0.1	3.7 ± 0.056	7.7 ± 0.079	1.8 ± 0.034	6.5 ± 0.07	5.1 ± 0.08	9.2 ± 0.086	2.2 ± 0.037	3.5 ± 0.061	4.3 ± 0.051	2.3 ± 0.042	5.8 ± 0.078	6.1 ± 0.07	5.5 ± 0.07	8.1 ± 0.073	1.0 ± 0.03	3.3 ± 0.05
Bacteria	9.9 ± 0.087	0.99 ± 0.024	5.4 ± 0.051	6.1 ± 0.074	3.9 ± 0.051	7.7 ± 0.068	2.1 ± 0.034	6.1 ± 0.06	4.8 ± 0.068	1e+01 ± 0.082	2.3 ± 0.033	3.6 ± 0.049	4.7 ± 0.05	3.6 ± 0.048	6.0 ± 0.065	6.0 ± 0.057	5.4 ± 0.051	7.2 ± 0.062	1.3 ± 0.03	2.9 ± 0.041
Eukaryota	7.6 ± 0.073	1.7 ± 0.035	5.5 ± 0.048	6.4 ± 0.072	3.9 ± 0.046	6.3 ± 0.071	2.4 ± 0.03	5.3 ± 0.052	5.8 ± 0.075	9.2 ± 0.078	2.2 ± 0.027	4.5 ± 0.051	5.3 ± 0.063	4.2 ± 0.059	5.6 ± 0.056	8.3 ± 0.077	5.6 ± 0.046	6.1 ± 0.052	1.2 ± 0.023	3.0 ± 0.038
Viruses	6.6 ± 0.45	2.1 ± 0.26	5.3 ± 0.32	5.7 ± 0.36	4.0 ± 0.29	6.6 ± 0.51	2.3 ± 0.24	5.3 ± 0.33	5.5 ± 0.48	9.1 ± 0.51	2.2 ± 0.22	4.3 ± 0.34	5.6 ± 0.39	4.2 ± 0.4	6.0 ± 0.41	7.3 ± 0.42	6.3 ± 0.38	6.5 ± 0.46	1.7 ± 0.15	3.4 ± 0.32

Values are in 10^{-2} format, ± indicates standard deviation via bootstrap 100

N-terminus

For all datasets the most frequent N-terminal amino acid is **methionine**.

For dataset:

- organism stats: 96.67%
- taxonomy stats: 97.96%
- pdb stats: 55.01%
- swissprot stats: 97.41%

Explanation:

Methionine is the initiator amino acid in protein synthesis. All proteins start with Met during translation, and it's retained when the second amino acid has a large side chain. According to the N-end rule, Met is stabilizing with a long half-life (>20h in yeast, ~30h in mammals).

Exercise 3

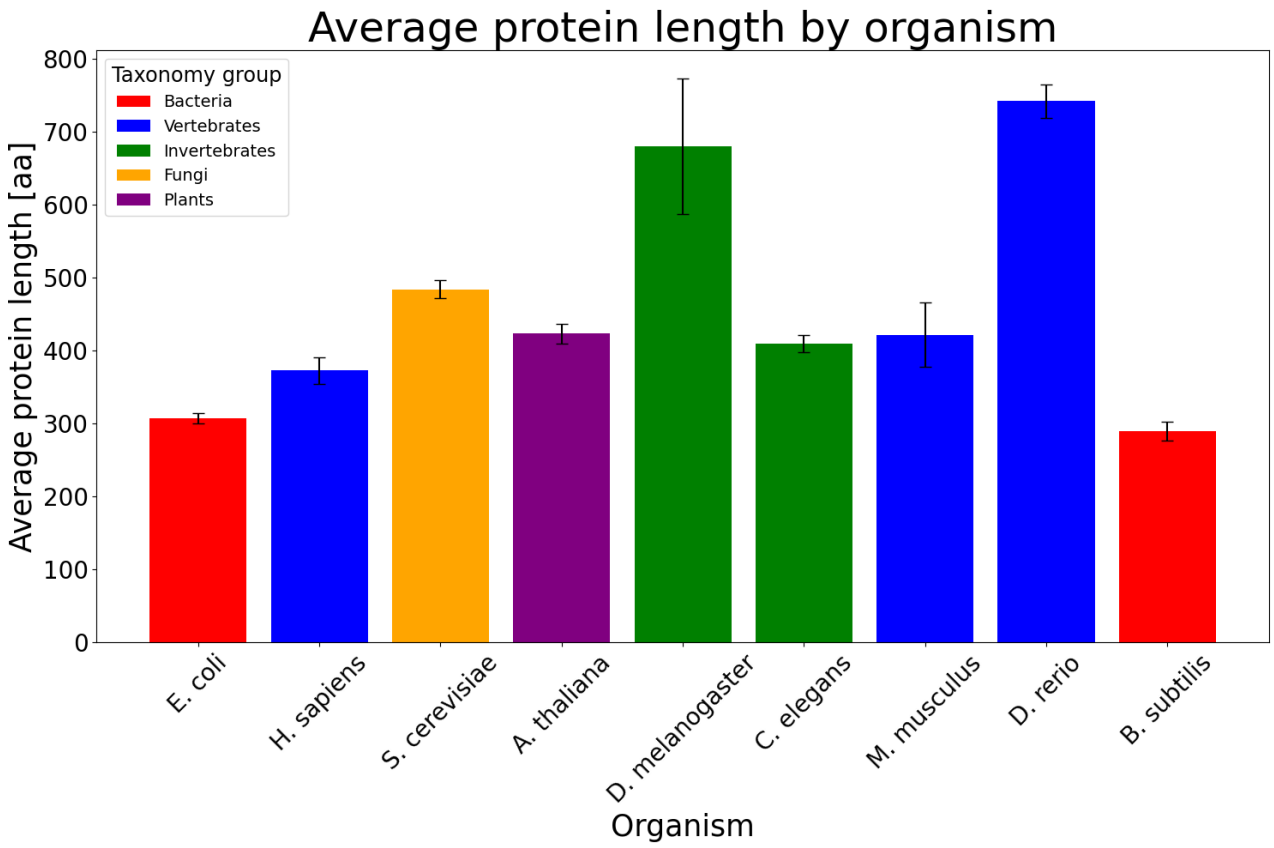


Table 5: Percantage content of amino acid by organism

Organism	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
E. coli	9.5%	1.2%	5.1%	5.8%	3.9%	7.4%	2.3%	6.0%	4.4%	1.1e+01%	2.8%	3.9%	4.4%	4.4%	5.5%	5.8%	5.4%	7.1%	1.5%	2.8%
H. sapiens	6.9%	2.2%	4.8%	7.2%	3.6%	6.5%	2.6%	4.3%	5.8%	9.9%	2.2%	3.6%	6.2%	4.8%	5.6%	8.4%	5.5%	6.0%	1.2%	2.6%
S. cerevisiae	5.5%	1.3%	5.8%	6.5%	4.4%	5.0%	2.2%	6.6%	7.3%	9.5%	2.1%	6.2%	4.4%	3.9%	4.4%	9.0%	5.9%	5.6%	1.0%	3.4%
A. thaliana	6.3%	1.9%	5.4%	6.8%	4.3%	6.3%	2.3%	5.3%	6.4%	9.6%	2.4%	4.4%	4.8%	3.5%	5.4%	9.2%	5.1%	6.6%	1.2%	2.8%
D. melanogaster	7.4%	1.9%	5.2%	6.6%	3.3%	6.2%	2.6%	4.8%	5.5%	8.7%	2.2%	4.7%	5.8%	5.4%	5.5%	8.6%	5.9%	5.9%	0.92%	2.8%
C. elegans	6.3%	2.1%	5.3%	6.5%	4.8%	5.4%	2.3%	6.2%	6.3%	8.6%	2.7%	4.9%	4.9%	4.1%	5.1%	8.1%	5.9%	6.2%	1.1%	3.2%
M. musculus	6.8%	2.3%	4.8%	7.0%	3.7%	6.4%	2.6%	4.4%	5.7%	1e+01%	2.2%	3.6%	6.2%	4.8%	5.6%	8.6%	5.4%	6.1%	1.2%	2.7%
D. rerio	6.3%	2.1%	5.3%	7.2%	3.4%	6.0%	2.7%	4.5%	6.0%	9.2%	2.3%	4.0%	5.8%	5.0%	5.5%	9.2%	5.8%	6.1%	1.0%	2.6%
B. subtilis	7.7%	0.79%	5.2%	7.3%	4.5%	6.9%	2.3%	7.4%	7.1%	9.7%	2.8%	4.0%	3.7%	3.8%	4.1%	6.3%	5.4%	6.7%	1.0%	3.5%

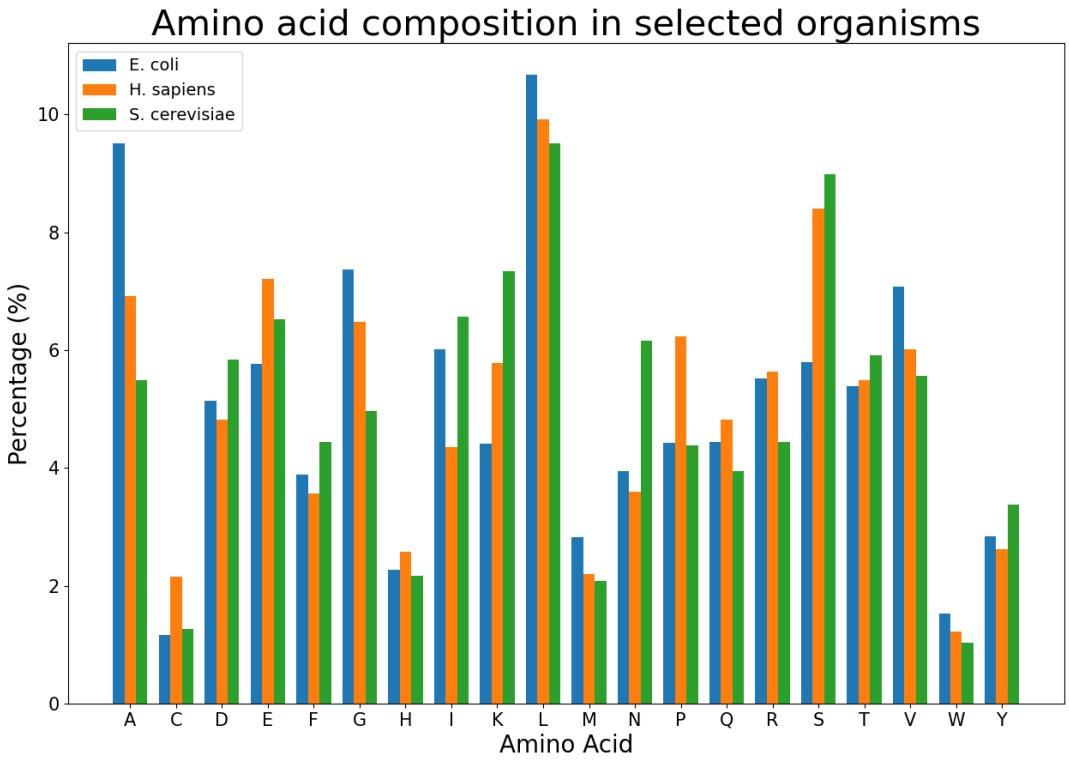


Table 6: PBD statistics

Average length [aa]	std	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
2.9e+02	4.8	7.9%	1.4%	5.4%	6.5%	3.9%	7.2%	2.5%	5.5%	6.0%	9.1%	2.3%	4.3%	4.7%	4.0%	5.3%	6.6%	5.7%	7.0%	1.3%	3.4%

Explanation why the statistics in PDB are different that those in proteomes from Uniprot:

Many PDB entries contain only domains or fragments of proteins rather than full-length sequences, as these are often selected for structural determination due to their biological, medical, or industrial relevance. Additionally, proteins in the PDB often undergo modifications (such as tags or mutations) to aid in crystallization or structure determination. Because proteins need to be isolated from the organism, they often undergo truncation, particularly at the ends, which may also lead to the removal of certain regions like the N-terminus. This can result in a smaller proportion of methionine (the typical N-terminal amino acid) in PDB compared to the full-length sequences found in UniProt(See N-terminus).

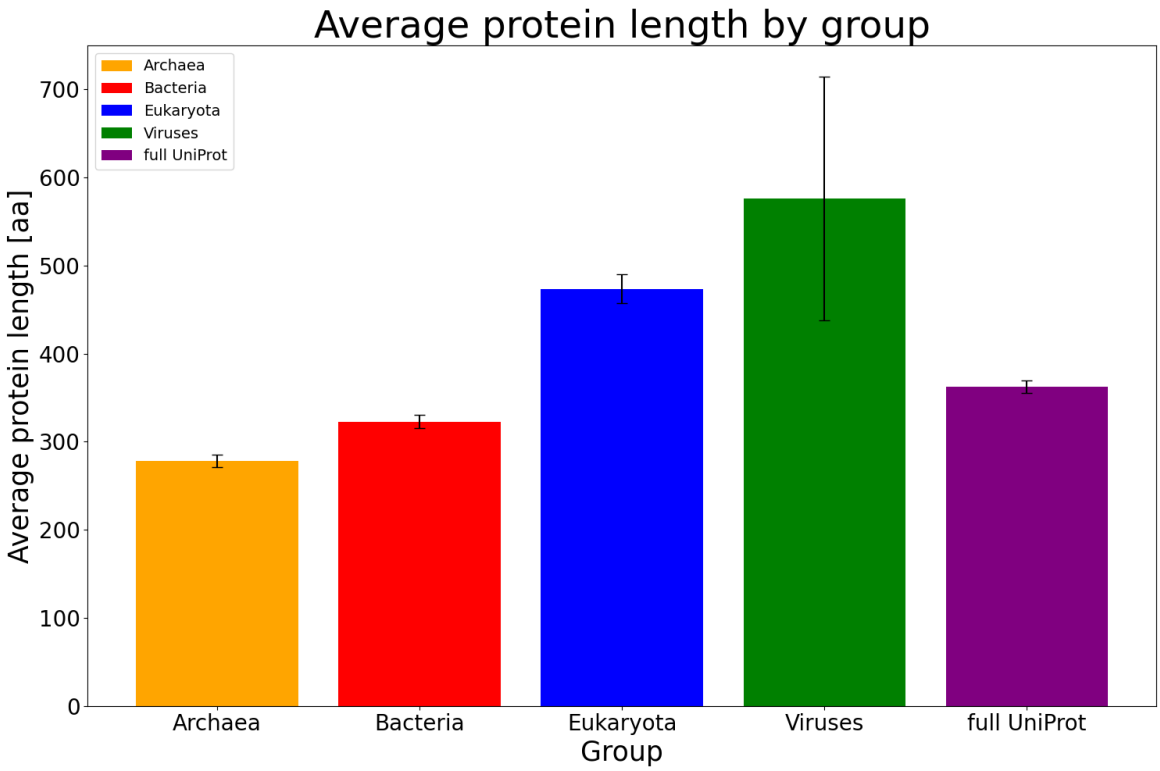
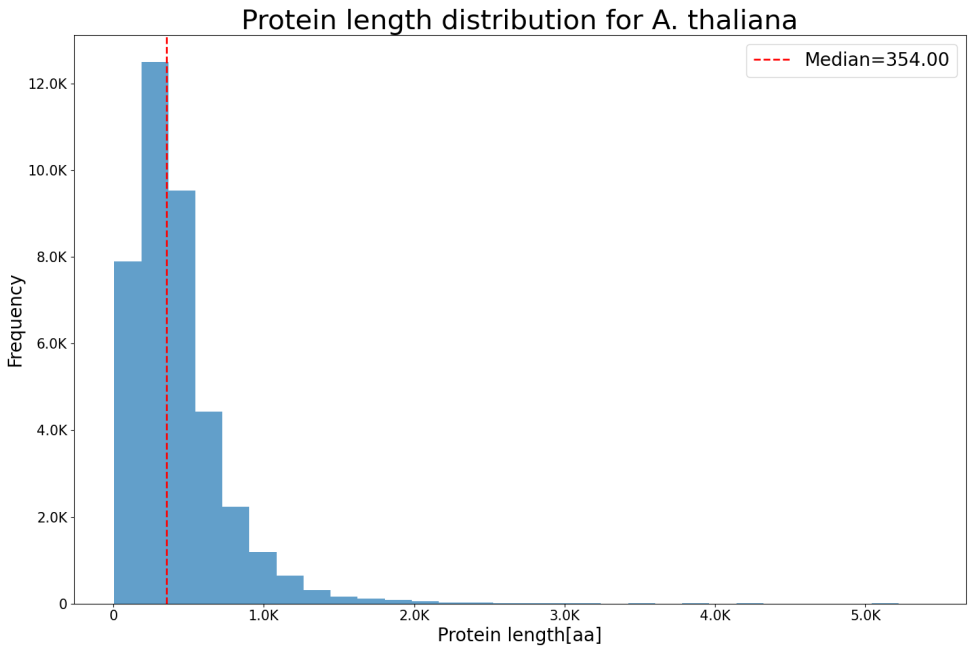
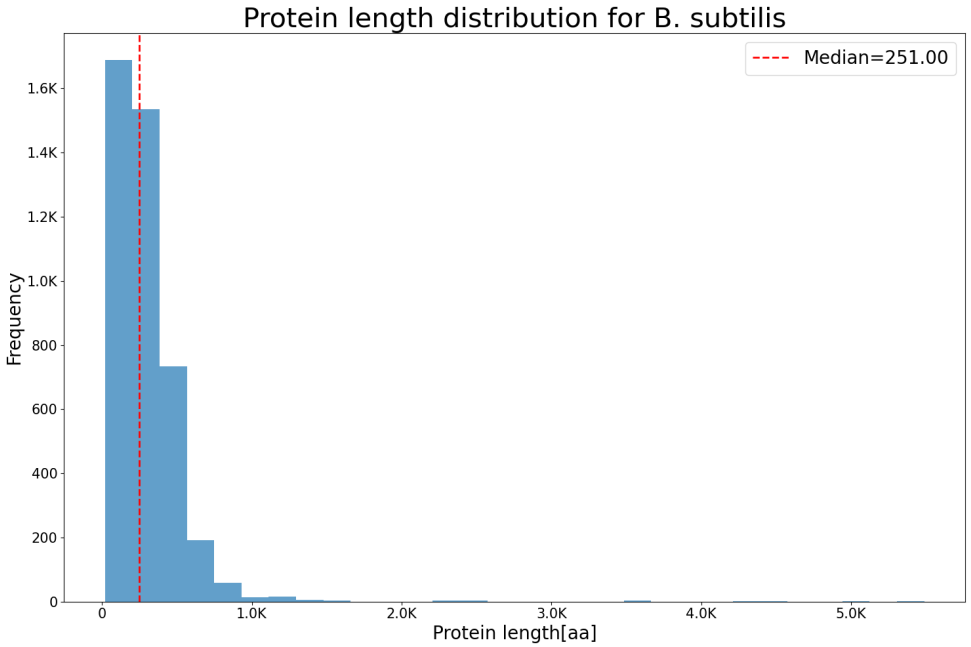


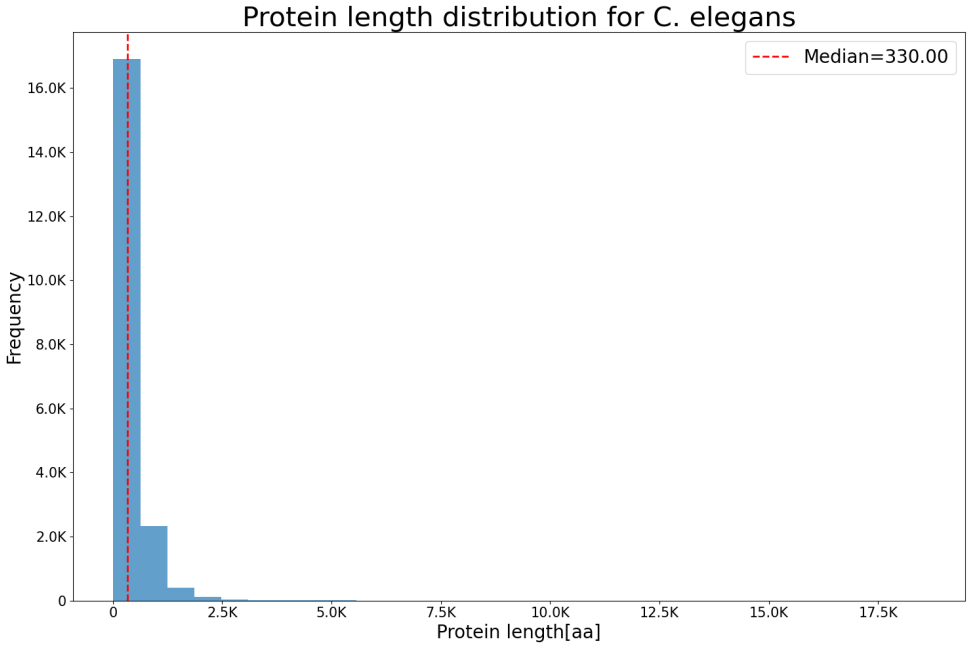
Table 7: Percantage content of amino acid by kingdom

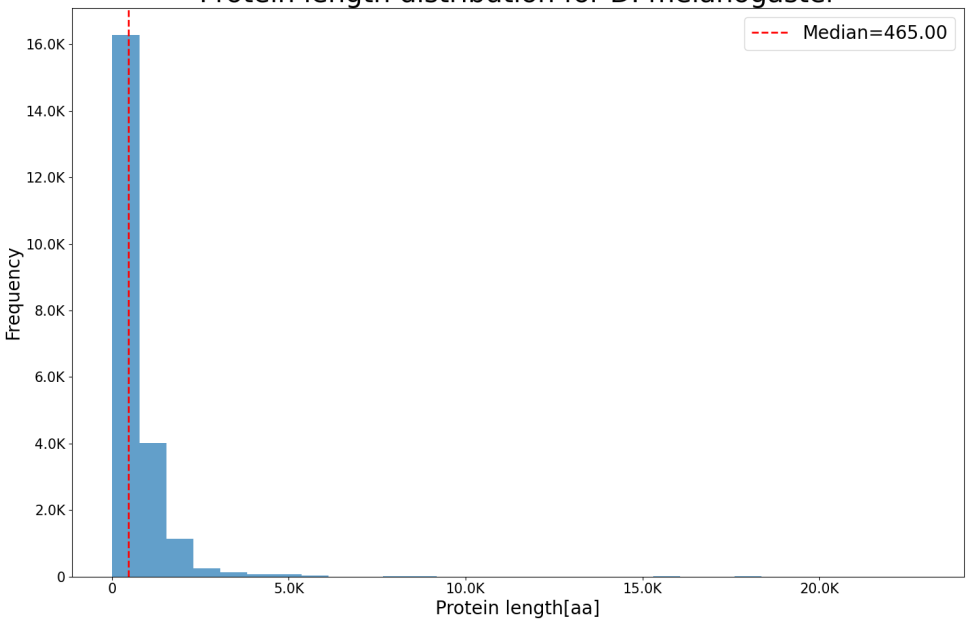
Group	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
full UniProt	8.3%	1.4%	5.5%	6.7%	3.9%	7.1%	2.3%	5.9%	5.8%	9.6%	2.4%	4.1%	4.7%	3.9%	5.5%	6.7%	5.4%	6.9%	1.1%	2.9%
Bacteria	9.9%	0.99%	5.4%	6.1%	3.9%	7.7%	2.1%	6.1%	4.8%	1e+01%	2.3%	3.6%	4.7%	3.6%	6.0%	6.0%	5.4%	7.2%	1.3%	2.9%
Viruses	6.6%	2.1%	5.3%	5.7%	4.0%	6.6%	2.3%	5.3%	5.5%	9.1%	2.2%	4.3%	5.6%	4.2%	6.0%	7.3%	6.3%	6.5%	1.7%	3.4%
Archaea	8.4%	0.89%	6.5%	8.1%	3.7%	7.7%	1.8%	6.5%	5.1%	9.2%	2.2%	3.5%	4.3%	2.3%	5.8%	6.1%	5.5%	8.1%	1.0%	3.3%
Eukaryota	7.6%	1.7%	5.5%	6.4%	3.9%	6.3%	2.4%	5.3%	5.8%	9.2%	2.2%	4.5%	5.3%	4.2%	5.6%	8.3%	5.6%	6.1%	1.2%	3.0%

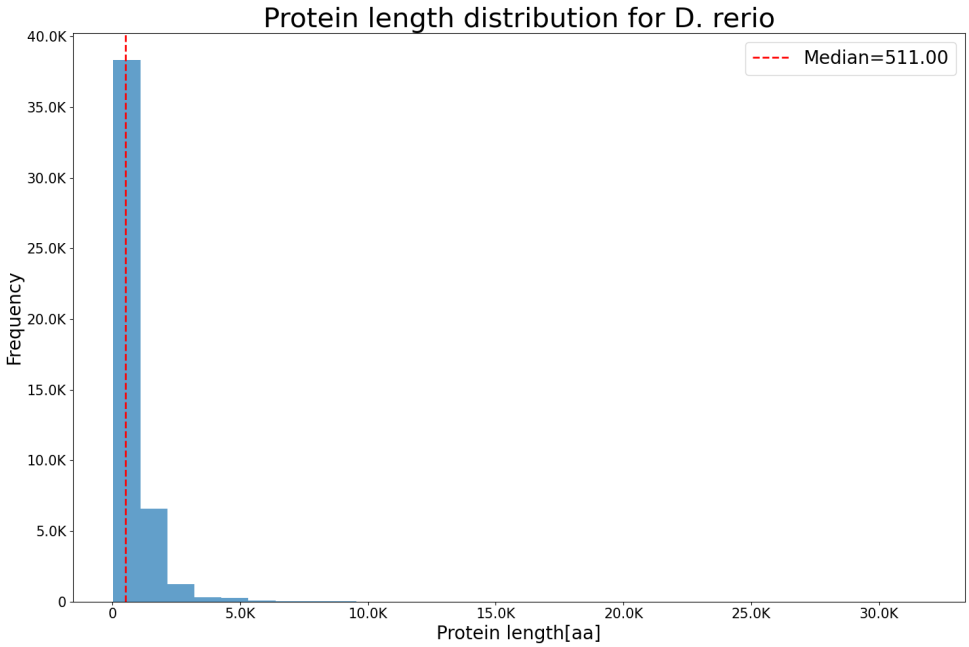
Histograms

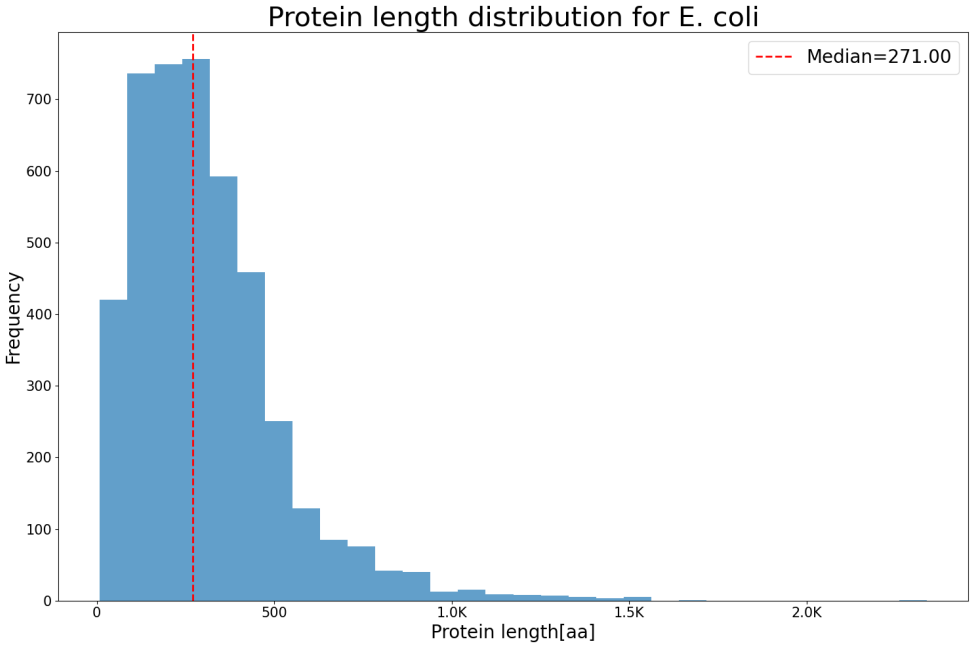


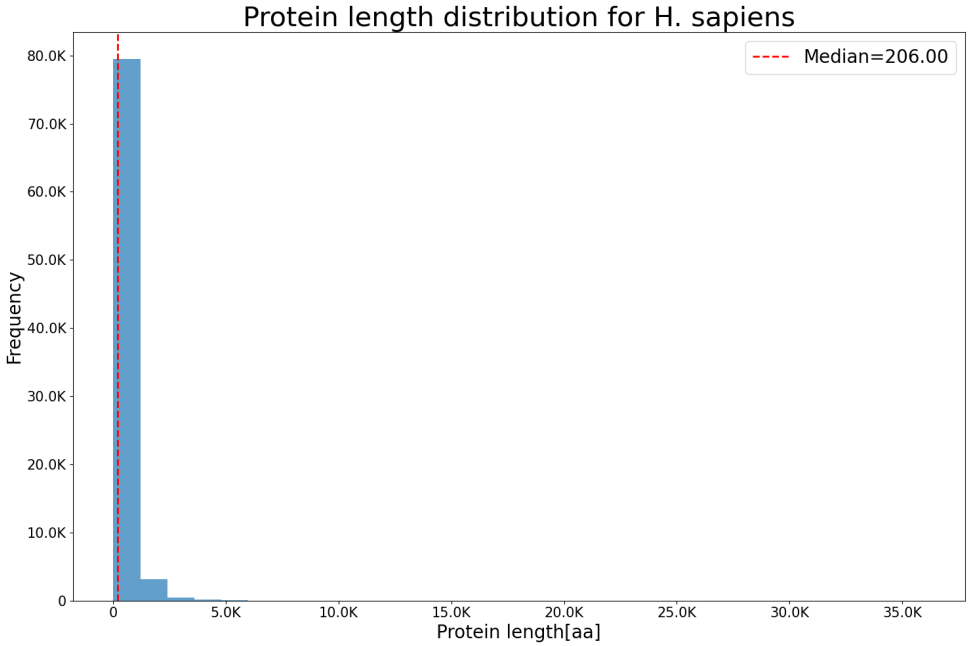

Protein length distribution for B. subtilis
Median=251.00
Frequency
Protein length[aa]
0
200
400
600
800
1.0K
1.2K
1.4K
1.6K
1.0K
2.0K
3.0K
4.0K
5.0K

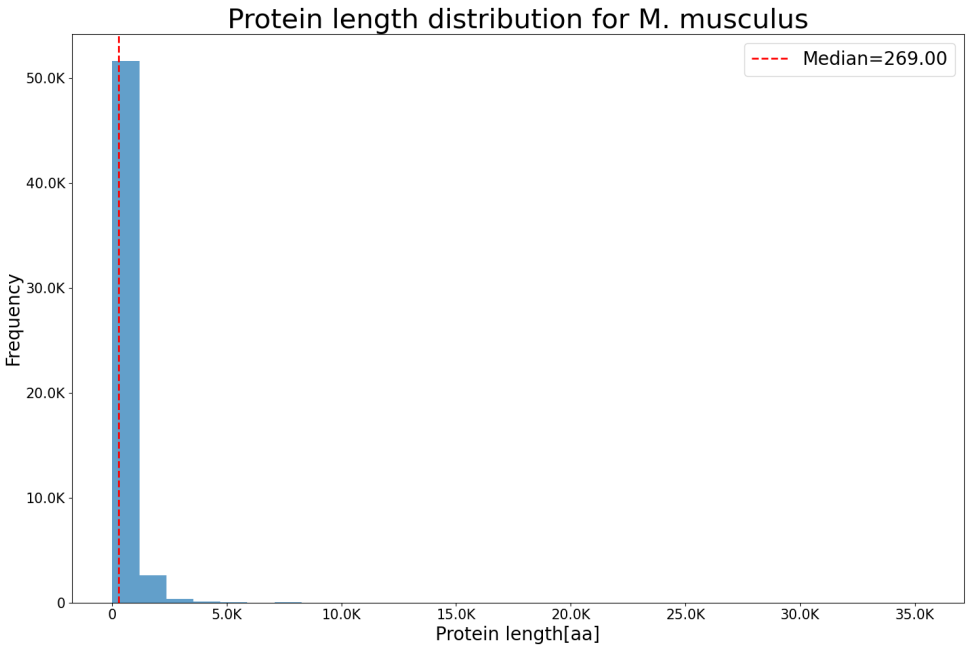

Protein length distribution for C. elegans
Median=330.00
Frequency
Protein length[aa]
0
2.0K
4.0K
6.0K
8.0K
10.0K
12.0K
14.0K
16.0K
2.5K
5.0K
7.5K
10.0K
12.5K
15.0K
17.5K

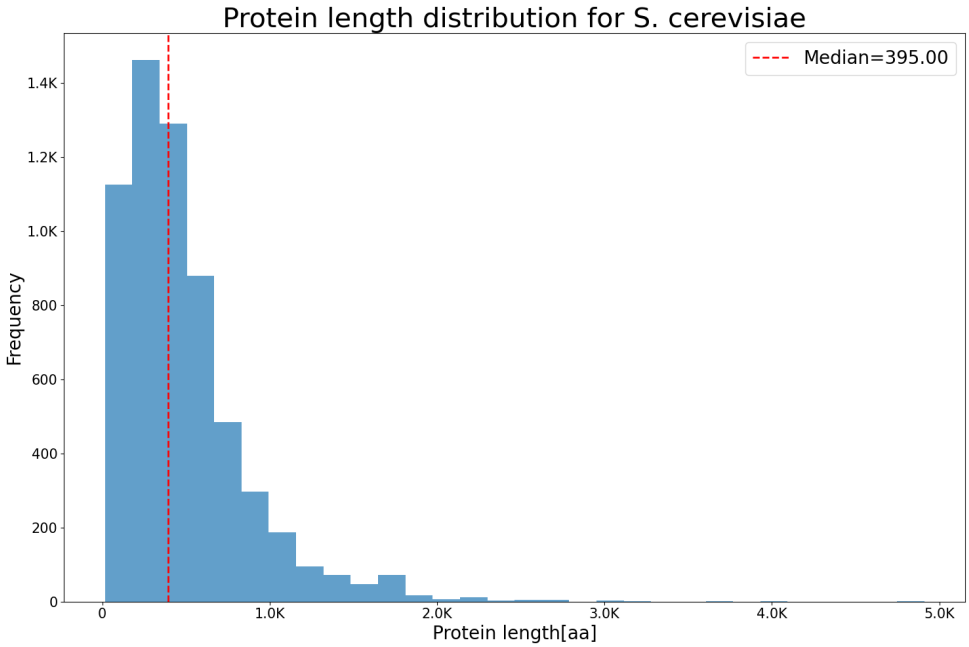

Protein length distribution for D. melanogaster
Median=465.00
Frequency
Protein length[aa]
0
2.0K
4.0K
6.0K
8.0K
10.0K
12.0K
14.0K
16.0K
5.0K
10.0K
15.0K
20.0K

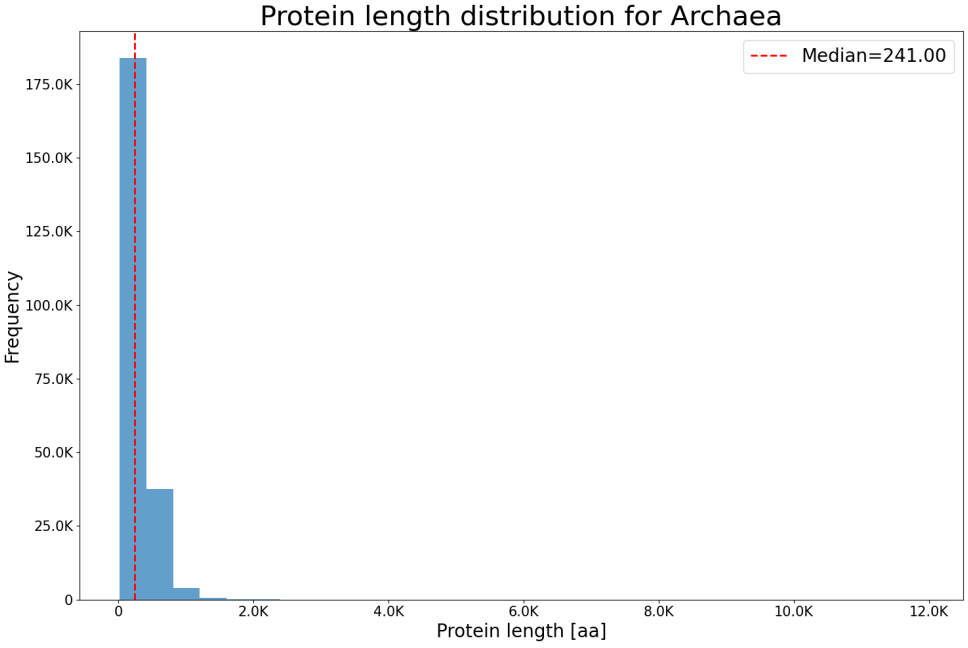

Protein length distribution for D. rerio
Median=511.00
Frequency
Protein length[aa]
0
5.0K
10.0K
15.0K
20.0K
25.0K
30.0K
35.0K
40.0K

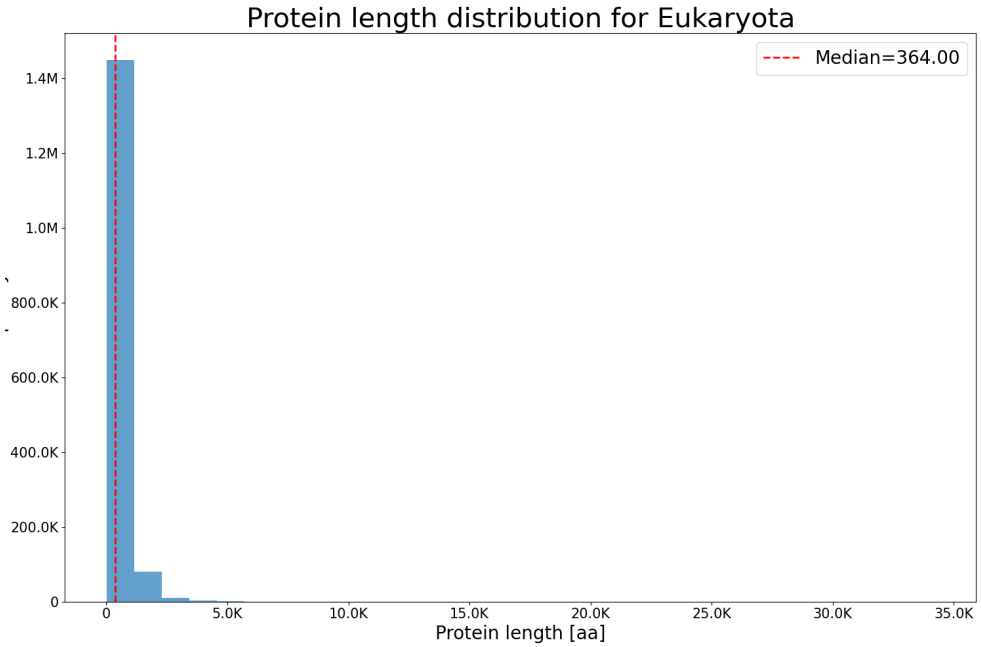

Protein length distribution for E. coli
Median=271.00
Frequency
Protein length[aa]
0
500
1.0K
1.5K
2.0K
100
200
300
400
600
700

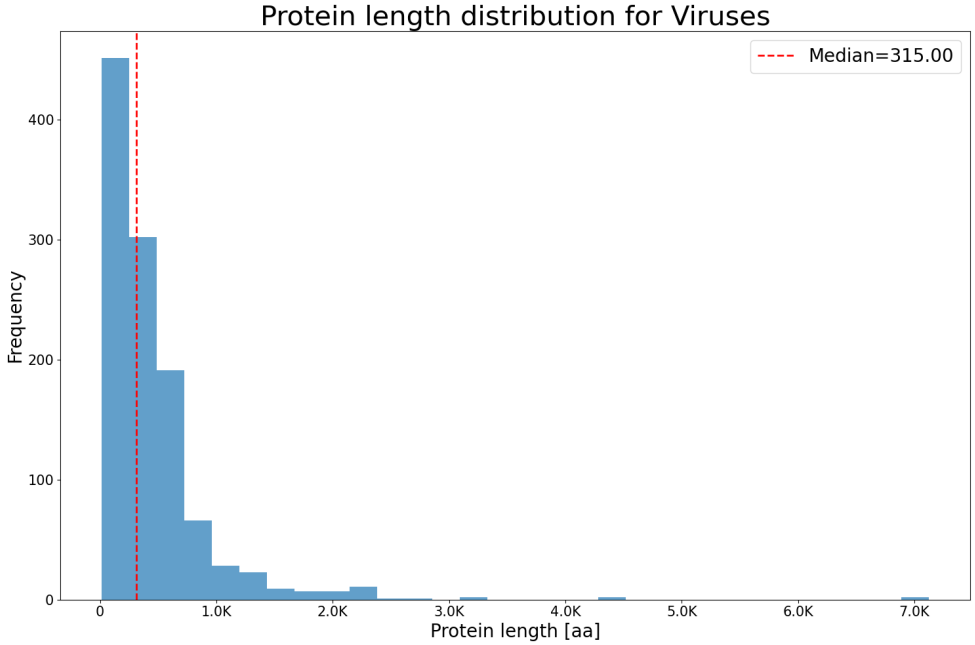

Protein length distribution for H. sapiens
Median=206.00
Frequency
Protein length[aa]
0
5.0K
10.0K
15.0K
20.0K
25.0K
30.0K
35.0K
10.0K
20.0K
30.0K
40.0K
50.0K
60.0K
70.0K
80.0K

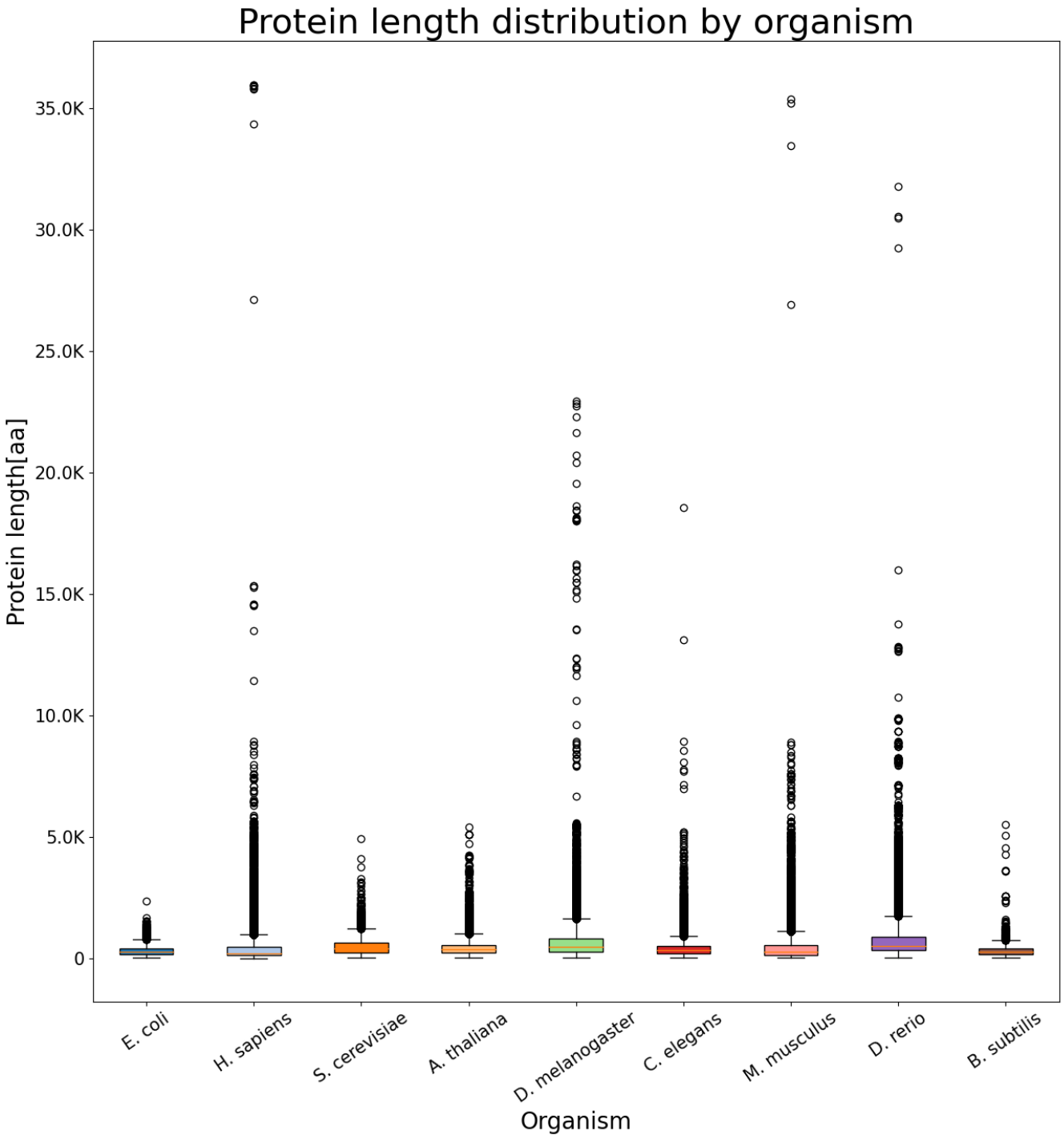

Protein length distribution for M. musculus
Median=269.00
Frequency
50.0K
40.0K
30.0K
20.0K
10.0K
0
0
5.0K
10.0K
15.0K
20.0K
25.0K
30.0K
35.0K
Protein length[aa]

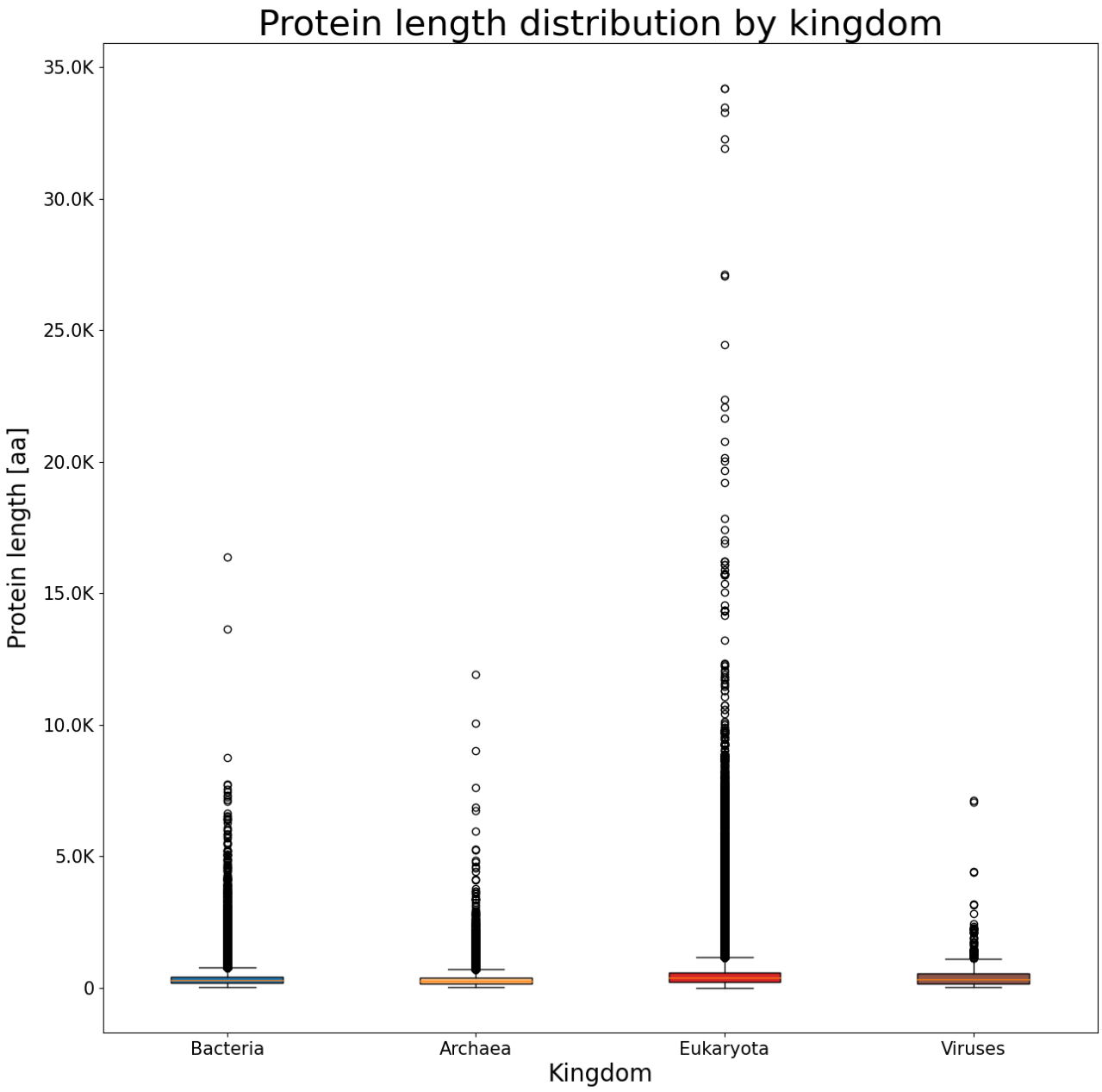

Protein length distribution for S. cerevisiae
Median=395.00
Frequency
1.4K
1.2K
1.0K
800
600
400
200
0
0
1.0K
2.0K
3.0K
4.0K
5.0K
Protein length[aa]


Protein length distribution for Archaea
Median=241.00
Frequency
175.0K
150.0K
125.0K
100.0K
75.0K
50.0K
25.0K
0
0
2.0K
4.0K
6.0K
8.0K
10.0K
12.0K
Protein length [aa]


Protein length distribution for Eukaryota
Median=364.00
1.4M
1.2M
1.0M
800.0K
600.0K
400.0K
200.0K
0
0
5.0K
10.0K
15.0K
20.0K
25.0K
30.0K
35.0K
Protein length [aa]


Protein length distribution for Viruses
Median=315.00
400
300
200
100
0
Frequency
0
1.0K
2.0K
3.0K
4.0K
5.0K
6.0K
7.0K
Protein length [aa]


Protein length distribution by organism
Protein length[aa]
35.0K
30.0K
25.0K
20.0K
15.0K
10.0K
5.0K
0
E. coli
H. sapiens
S. cerevisiae
A. thaliana
D. melanogaster
C. elegans
M. musculus
D. rerio
B. subtilis
Organism



Which is better: median or arithmetic mean?

Arithmetic Mean

1. Pros:
    1. Considers all data points in the calculation
    2. Widely used in scientific literature, facilitating comparisons
    3. Works well for normally distributed data
2. Cons:
    1. Highly sensitive to outliers
    2. Can be misleading in skewed distributions

3. May not represent the "typical" protein in non-normal distributions

Median

1. Pros:
    1. Robust against outliers, which are common in protein length data
    2. Better represents the "typical" protein in skewed distributions
    3. Divides the dataset exactly in half
    4. No assumptions about underlying distribution
2. Cons:
    1. Ignores the magnitude of extreme values
    2. Less sensitive to sample variation
    3. Less mathematically tractable for some statistical analyses

Conclusions

Protein length distributions are typically right-skewed (with a long tail toward longer proteins) because most proteins are of moderate length with fewer very large proteins. In these cases, the median often provides a better representation of the "typical" protein in an organism or kingdom.

For example, in our histograms, we can observe that the mean is often pulled higher than the median due to the influence of long proteins. This effect is particularly pronounced in eukaryotes, which have many large multi-domain proteins alongside smaller ones.

For comprehensive analysis, reporting both statistics is ideal: the median to represent the typical protein and the mean to account for the total protein mass distribution.