Big Data: report

Paulina Kucharewicz

2023-03-31

Part 1

Task 1

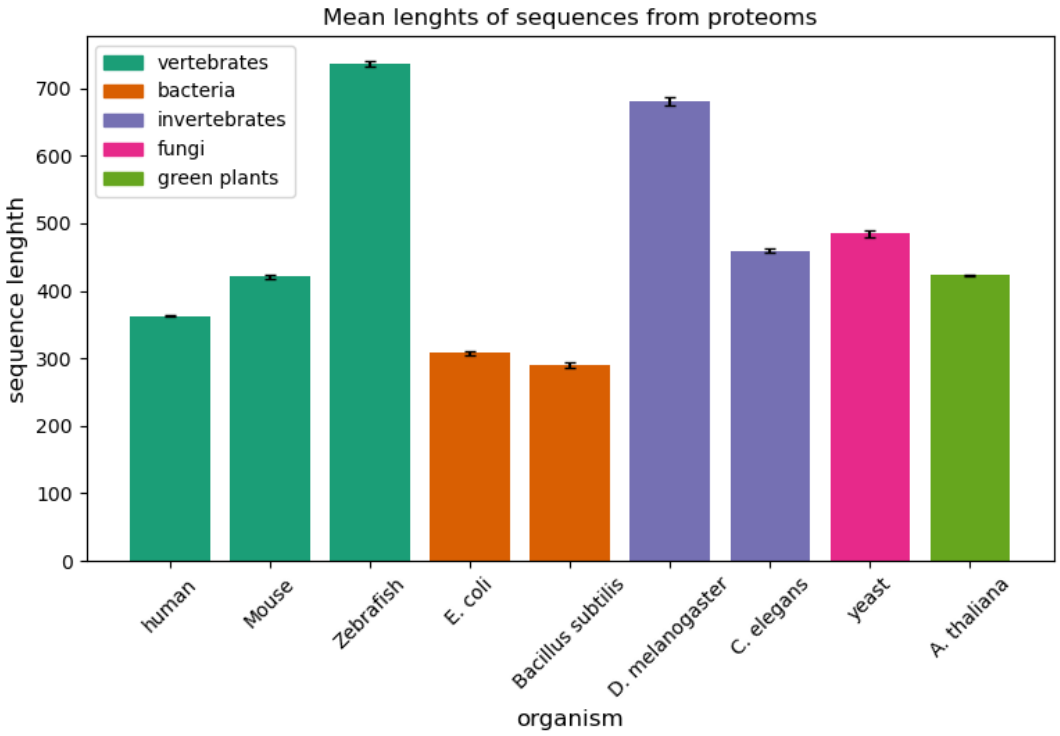

Figure 1: Comparison of average protein length between selected organisms.

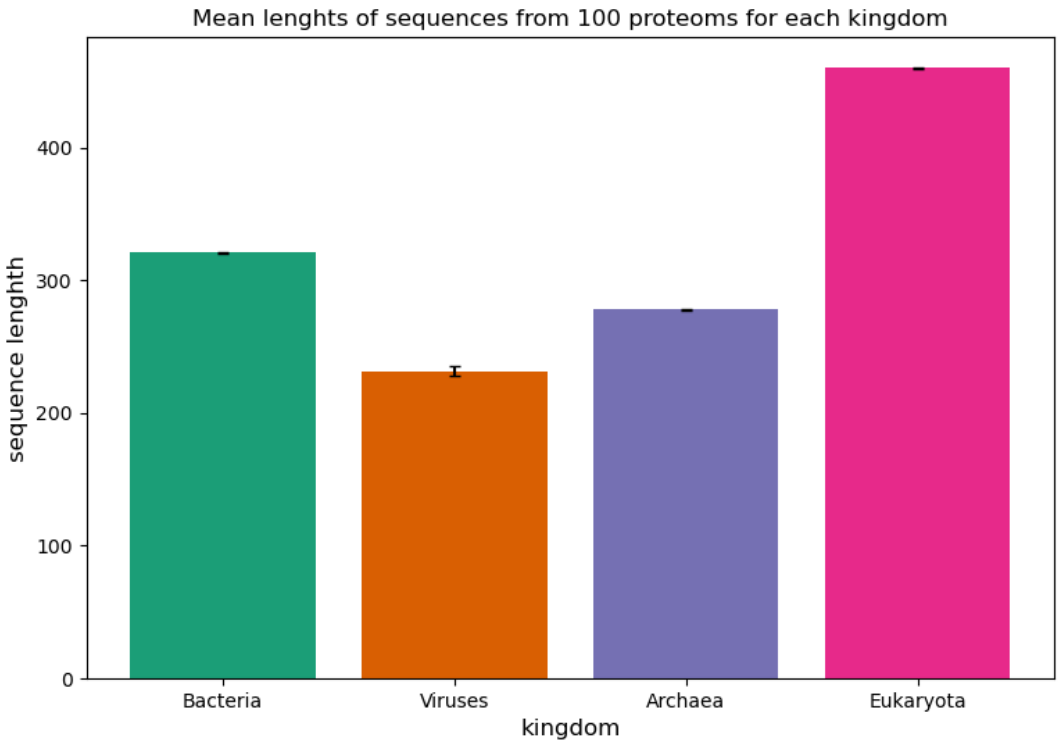

Figure 2: Comparison of average protein length between kingdoms.

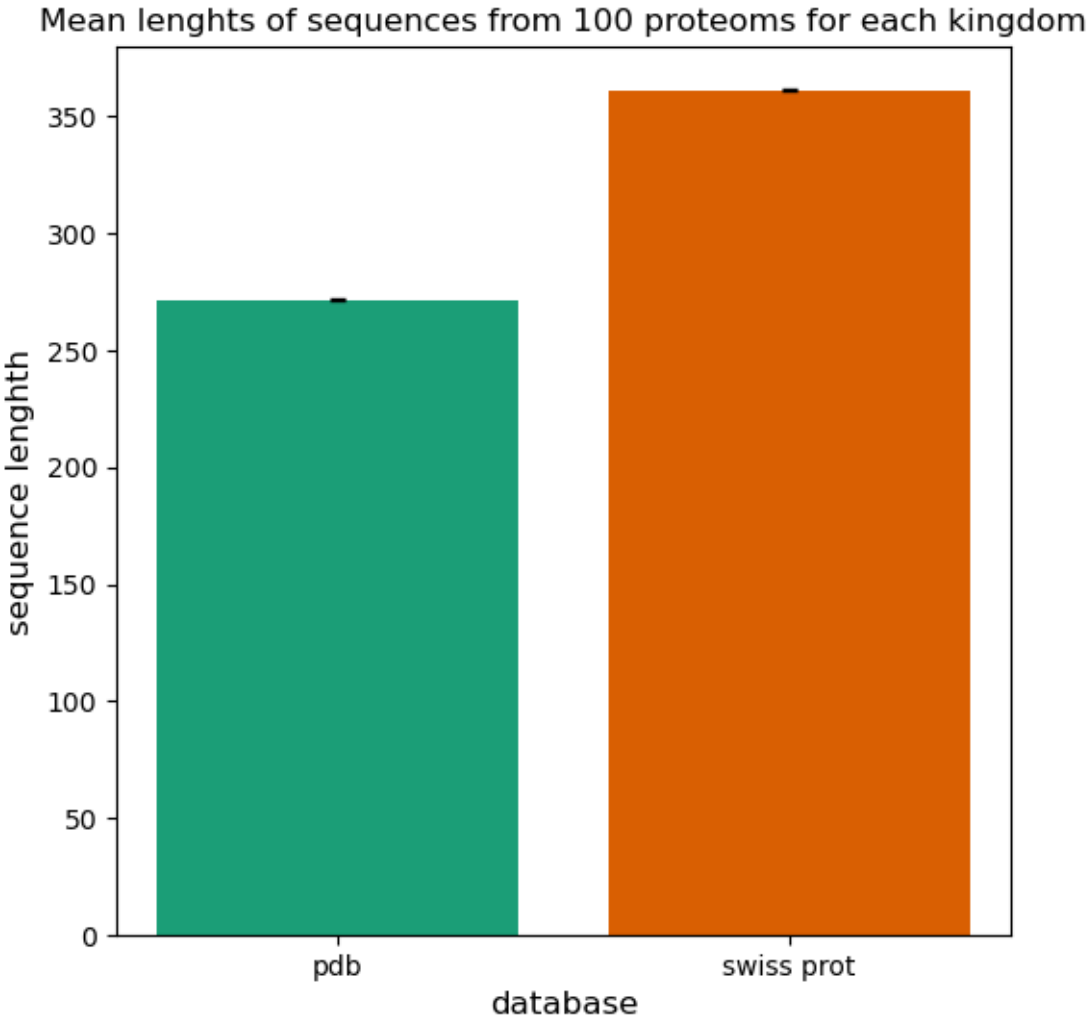

Figure 3: Comparison of average protein length between PDB and Uniprot.

Task 2

Table 1: Average aa content in protein sequences for selected organisms.

aminoacid	E._coli	B._subtilis	human	yeast	A._thaliana	D._melanog.	C._elegans	Mouse	Zebrafish
A	9.3	7.3	7.1	5.7	6.3	7.4	6.4	6.9	6.3
C	1.3	0.9	2.3	1.4	2	2	2.2	2.4	2.3
D	5	5.1	4.6	5.6	5.2	5.1	5.1	4.6	5.2
E	5.8	7.3	6.8	6.3	6.6	6.2	6.3	6.7	6.9
F	3.9	4.6	3.7	4.5	4.4	3.7	4.9	3.8	3.7
G	7	6.6	6.7	5.1	6.4	6.3	5.5	6.5	6
H	2.3	2.3	2.5	2.2	2.3	2.6	2.3	2.6	2.7
I	6.2	7.5	4.2	6.5	5.4	5	6.2	4.3	4.6

aminoacid	E._coli	B._subtilis	human	yeast	A._thaliana	D._melanog.	C._elegans	Mouse	Zebrafish
K	4.7	7.5	5.7	7.5	6.5	5.7	6.3	5.6	6.1
L	10.6	9.6	10.1	9.5	9.4	9.1	8.5	10.2	9.3
M	3.1	3	2.5	2.2	2.6	2.5	2.8	2.6	2.5
N	3.9	4	3.4	5.8	4.3	4.7	4.8	3.5	4
P	4.3	3.5	6.2	4.3	4.8	5.4	4.9	6	5.5
Q	4.4	3.9	4.7	4	3.4	5	4.1	4.7	4.7
R	5.6	4.1	5.9	4.7	5.5	5.6	5.2	5.8	5.6
S	5.8	6.2	8.1	8.6	9.1	8.1	8	8.3	8.8
T	5.3	5.3	5.2	5.8	5.1	5.6	5.8	5.3	5.7
V	7.1	6.7	6	5.7	6.7	6	6.2	6.1	6.2
W	1.5	1	1.4	1.1	1.2	1	1.1	1.3	1.1
Y	2.8	3.5	2.6	3.4	2.9	3.1	3.3	2.7	2.8

Table 2: Average protein length content for selected organisms.

	E._coli	B._subtilis	human	yeast	A._thaliana	D._melanog.	C._elegans	Mouse	Zebrafish
mean seq. length	307.619	289.775	362.953	484.722	423.182	681.09	459.093	420.928	736.859

Table 3: Average aa content in protein sequences for kingdoms.

aminoacid	swiss_prot	Bacteria	Viruses	Archaea	Eukaryota
A	8.4	9.9	7.7	8.8	7.2
C	1.5	1	1.5	1	2
D	5.3	5.5	6	6.9	5.2
E	6.6	6.3	7.1	8.2	6.3
F	3.9	4	3.9	3.6	4
G	7.2	7.5	6.6	7.7	6.3
H	2.2	2	2.1	1.9	2.5
I	6.1	6.1	5.8	6.1	5.2
K	6.2	5	6.5	4.4	5.8
L	9.6	10	8.3	9.2	9.4
M	2.6	2.5	2.8	2.3	2.4
N	3.9	3.6	4.4	3.2	4.2
P	4.5	4.5	4.2	4.3	5.3
Q	3.8	3.5	3.4	2.5	4.1
R	5.8	5.9	5.8	6	5.7
S	6.3	5.7	6.1	6	8.2
T	5.3	5.4	5.9	5.7	5.5
V	7.1	7.2	6.8	8.2	6.3
W	1.1	1.3	1.6	1.1	1.3
Y	2.9	2.9	3.5	3.1	2.9

Table 4: Average protein length content for kingdoms.

	swiss_prot	Bacteria	Viruses	Archaea	Eukaryota
mean seq. length	361.426	321.024	231.65	277.976	460.139

Task 3

Most frequent at N-terminus for all datasets was methionine (M). Aminoacid at N-terminus influences protein's stability and proteins with methionine at the N-terminus are more stable than most (or all, depending on organism) of the other aminoacids.

Part 2

Task a

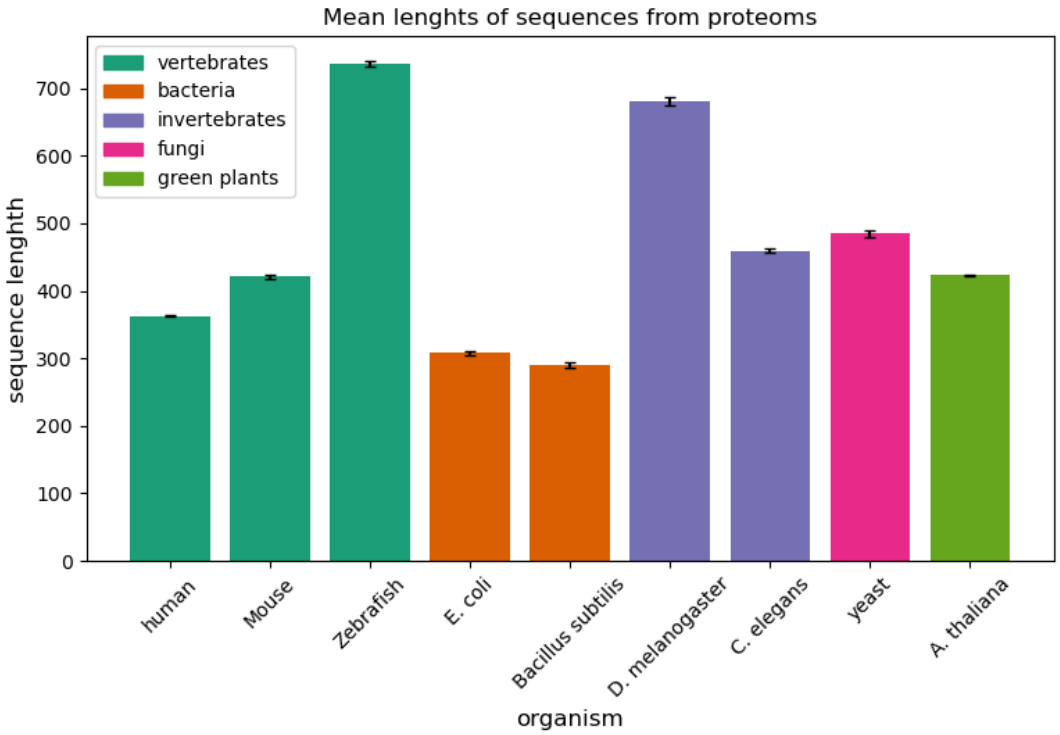

Figure 4: Comparison of average protein length between selected organisms.

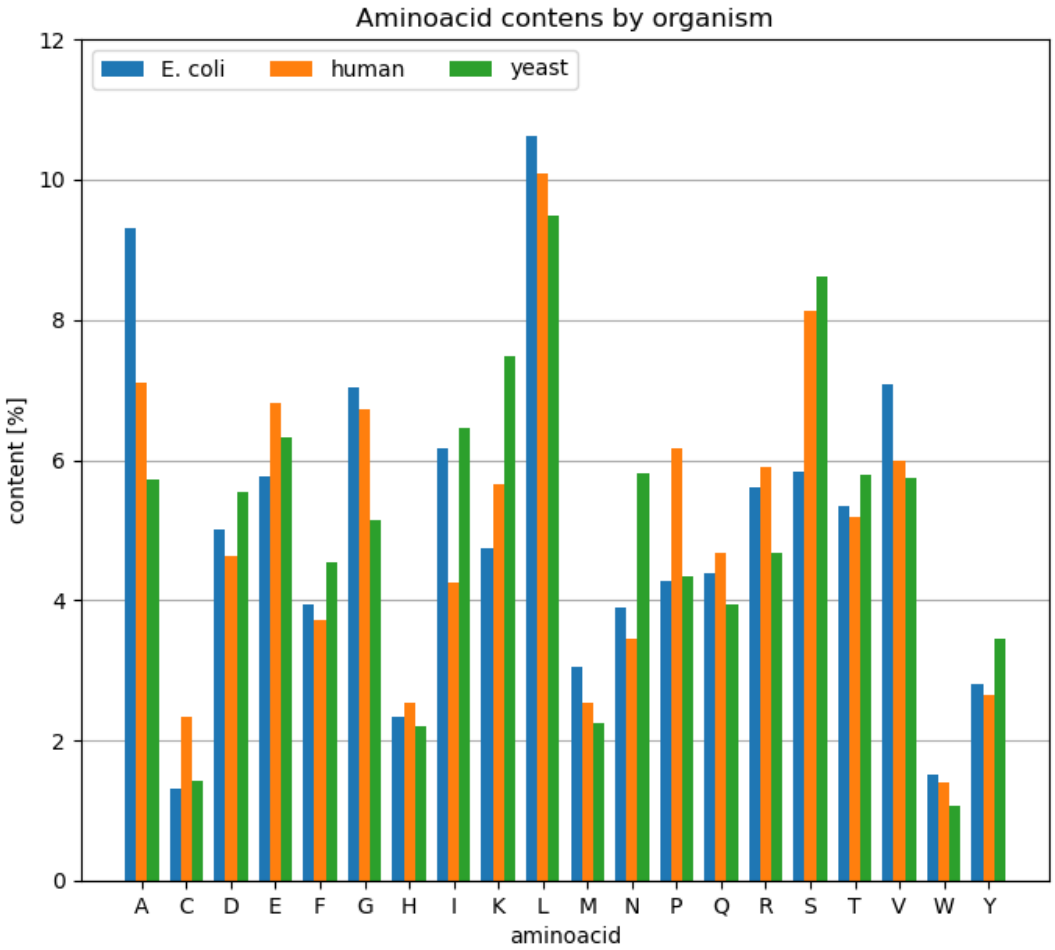

Figure 5: Comparison of percentage content of all amino acids for E.coli,human, yeast.

Task b

Table 5: Average aa content in protein sequences for PDB.

aminoacid	pdb
A	9
C	2.8
D	4.9
E	6
F	3.5
G	8.4
H	2.4
I	5.1
K	6.1
L	8.3

aminoacid	pdb
M	2.2
N	3.8
P	4.3
Q	3.6
R	5.4
S	6
T	6.1
V	6.6
W	1.2
Y	3.1

Table 6: Average protein length content for PDB.

	pdb
mean seq. length	271.744

Statistics from all PDB records are impacted by significant amount of molecules that are not proteins but DNA and RNA - their sequences are much shorter and skew mean sequence length and consist of nucleic acids represented by letters C, G, A, T which are counted as aminoacids and skew those statistics (higher frequency of those aa than could be expected).

Task c

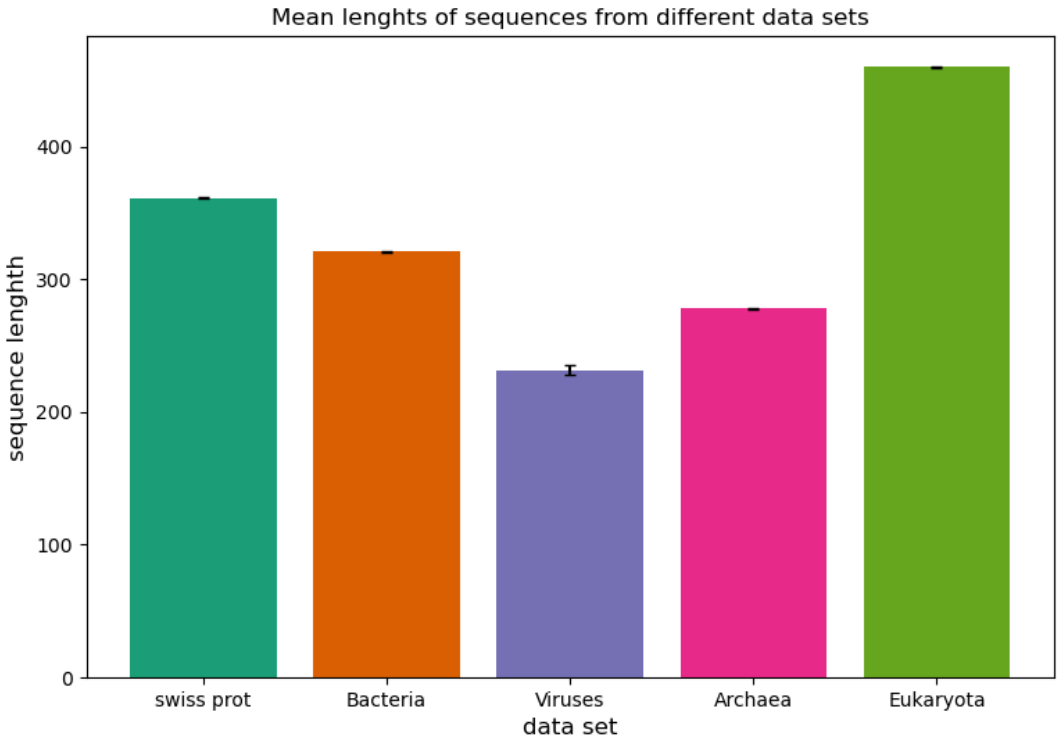

Figure 6: Comparison of average protein length between kingdoms and full uniprot.

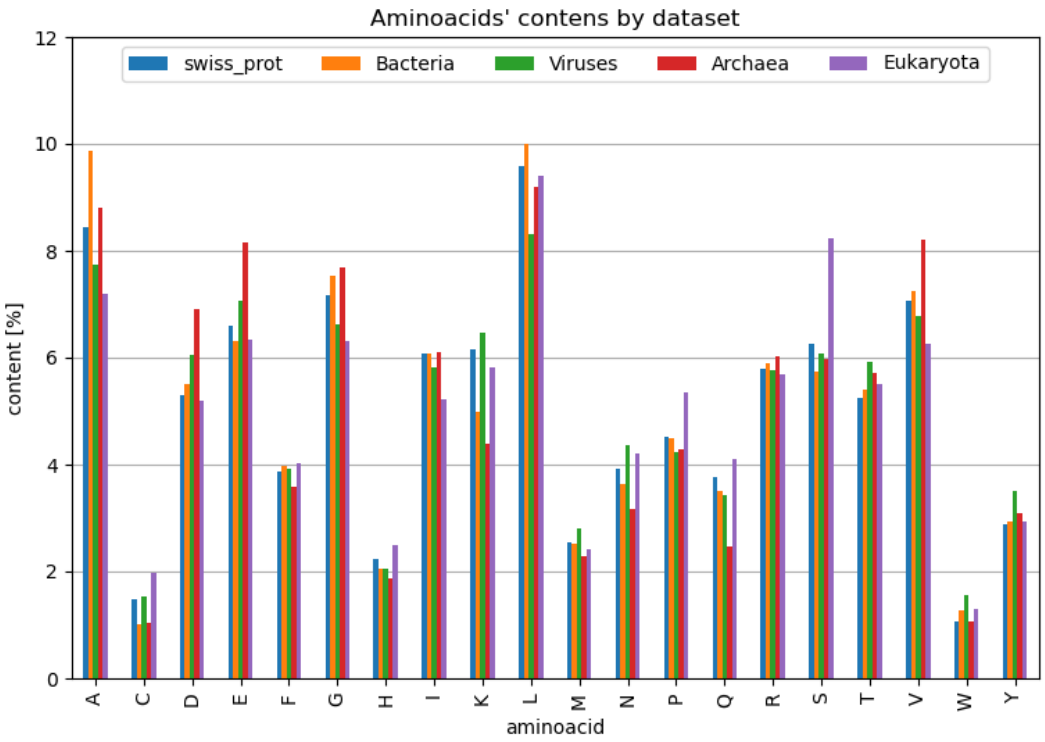

Figure 7: Comparison of percentage content of all amino acids between kingdoms and full uniprot.

Task d

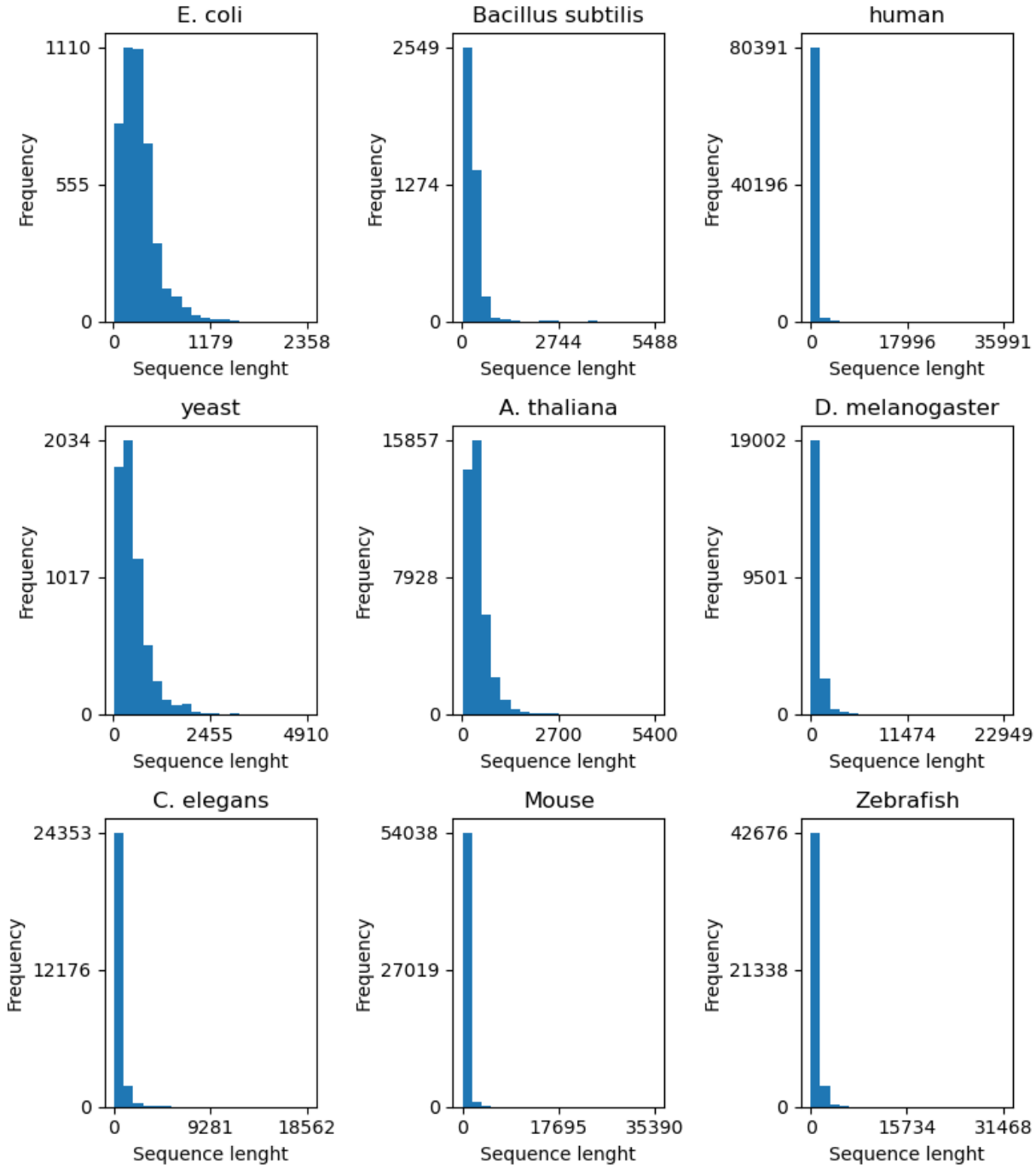

Figure 8: Comparison of protein length distribution for selected organisms.

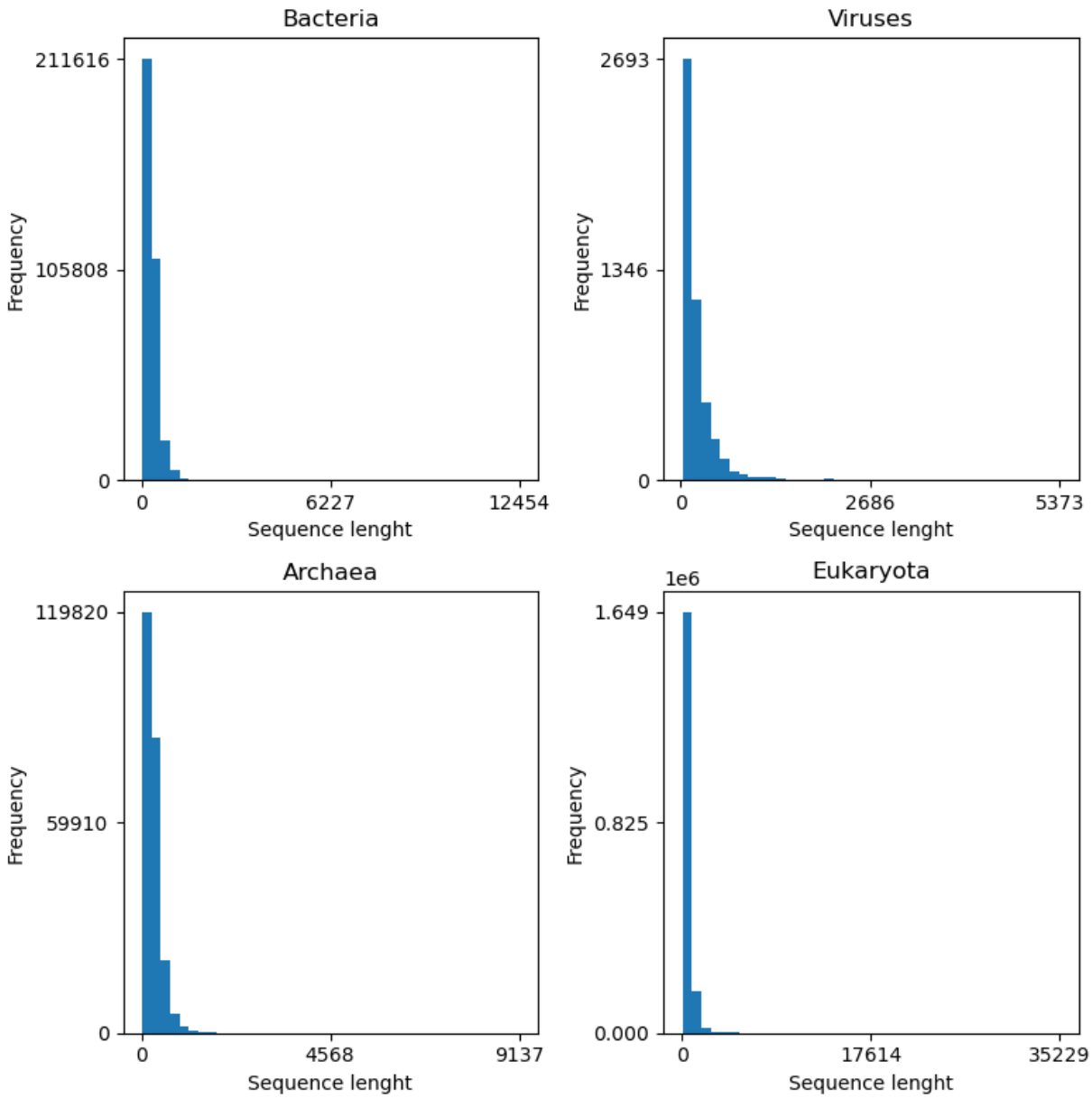

Figure 9: Comparison of protein length distribution for kingdoms.

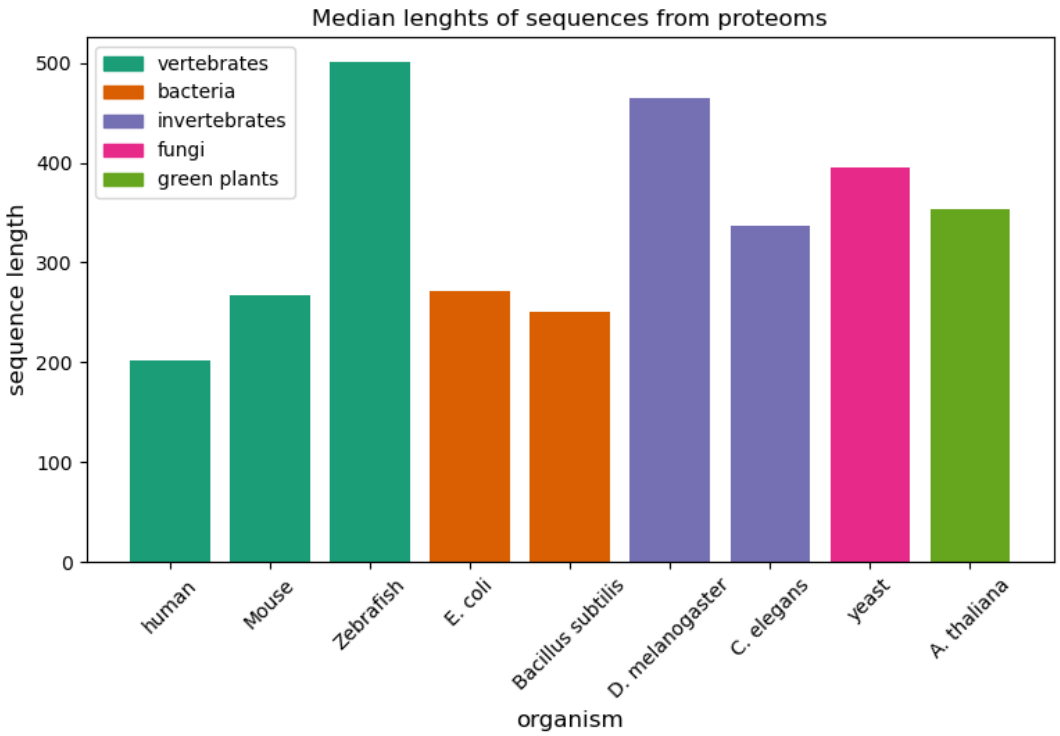

Figure 10: Comparison of median protein length between selected organisms.

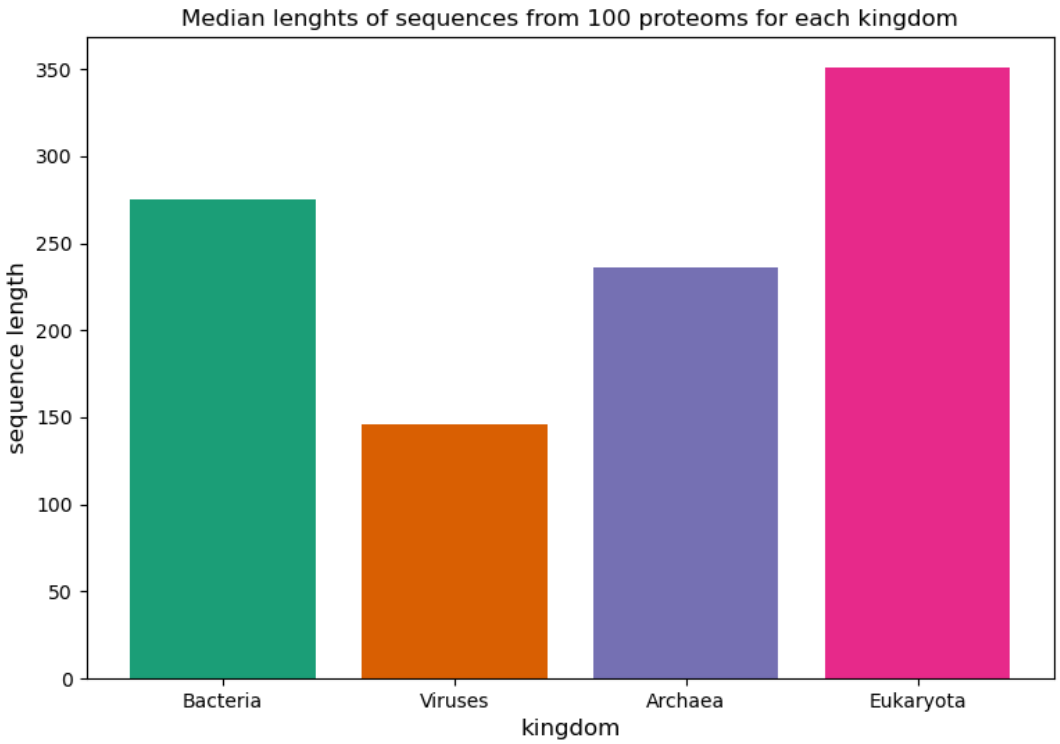

Figure 11: Comparison of median protein length between kingdoms.

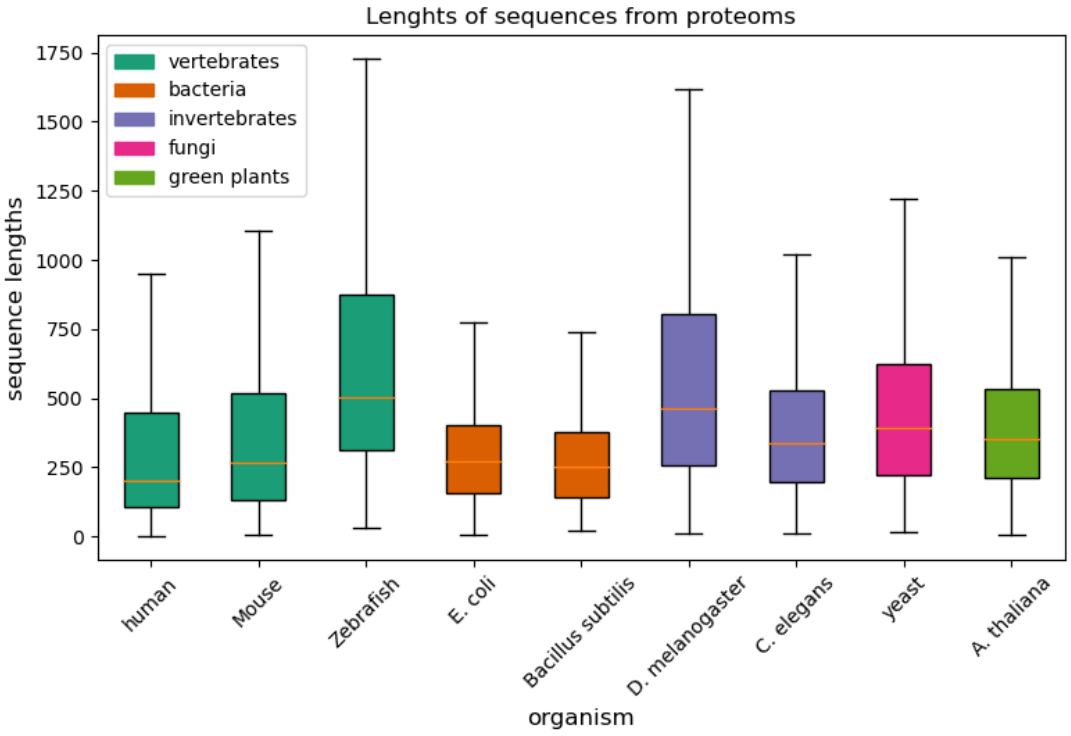

Figure 12: Comparison of protein length distribution for selected organisms.

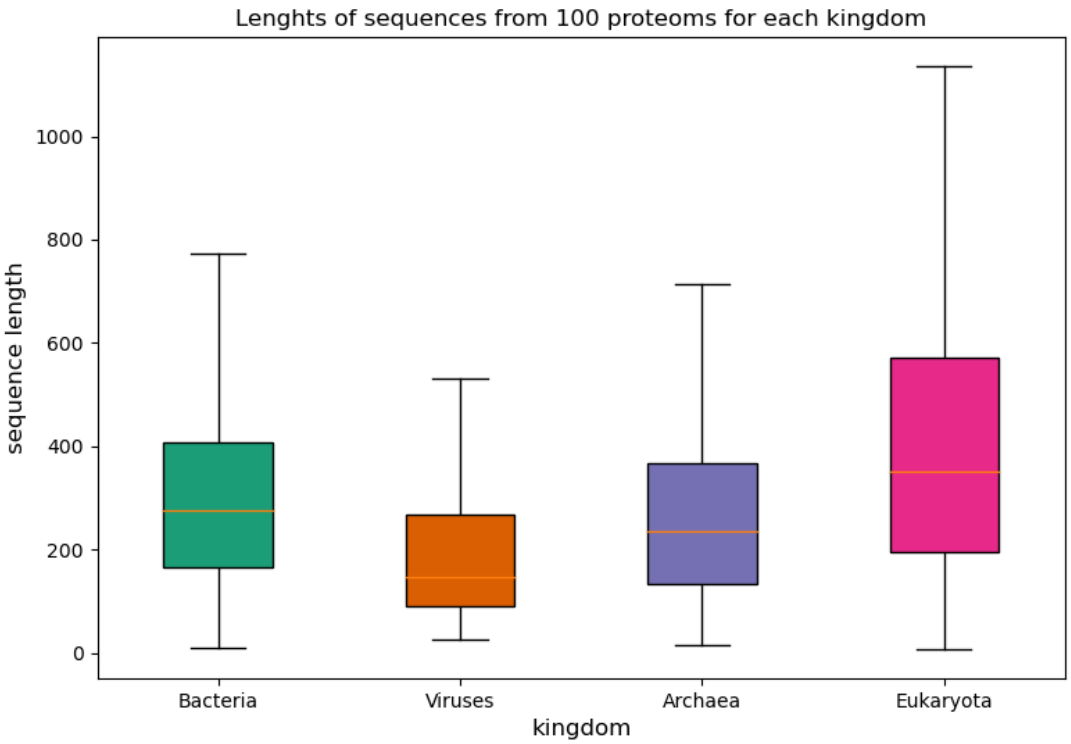

Figure 13: Comparison of protein length distribution for kingdoms.

Discuss which is better: median or arithmetic mean (prons and cons)?

Mean can be skewed if data distribution is not symmetric. In that case median can describe data better as mean value can be

If data distribution is not symmetric arithmetic mean can be skewed by outliers. If that's the case, median can be more descriptive of central tendency of the data distribution. Mean has an advantage of taking into account every value in the set.