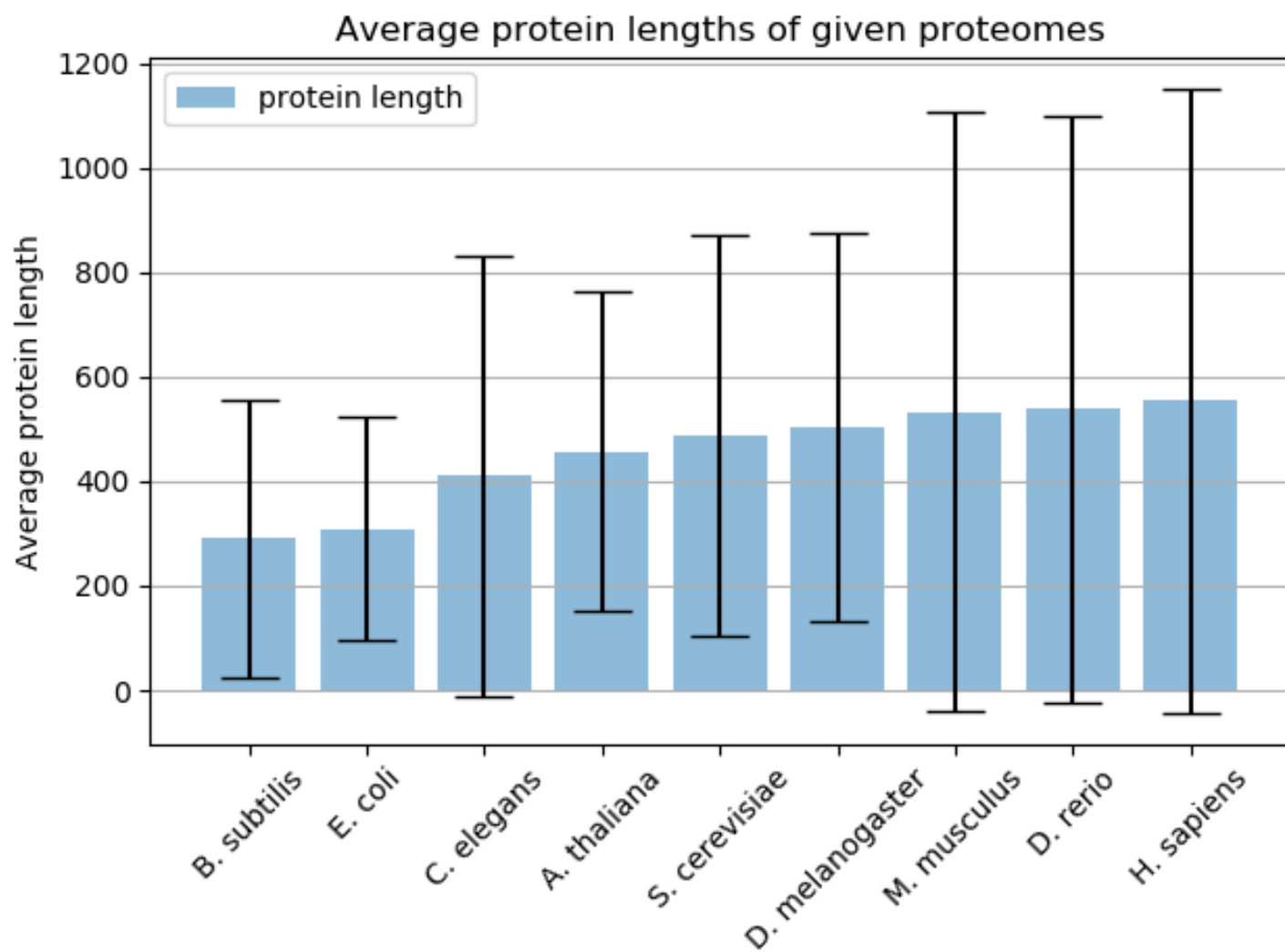
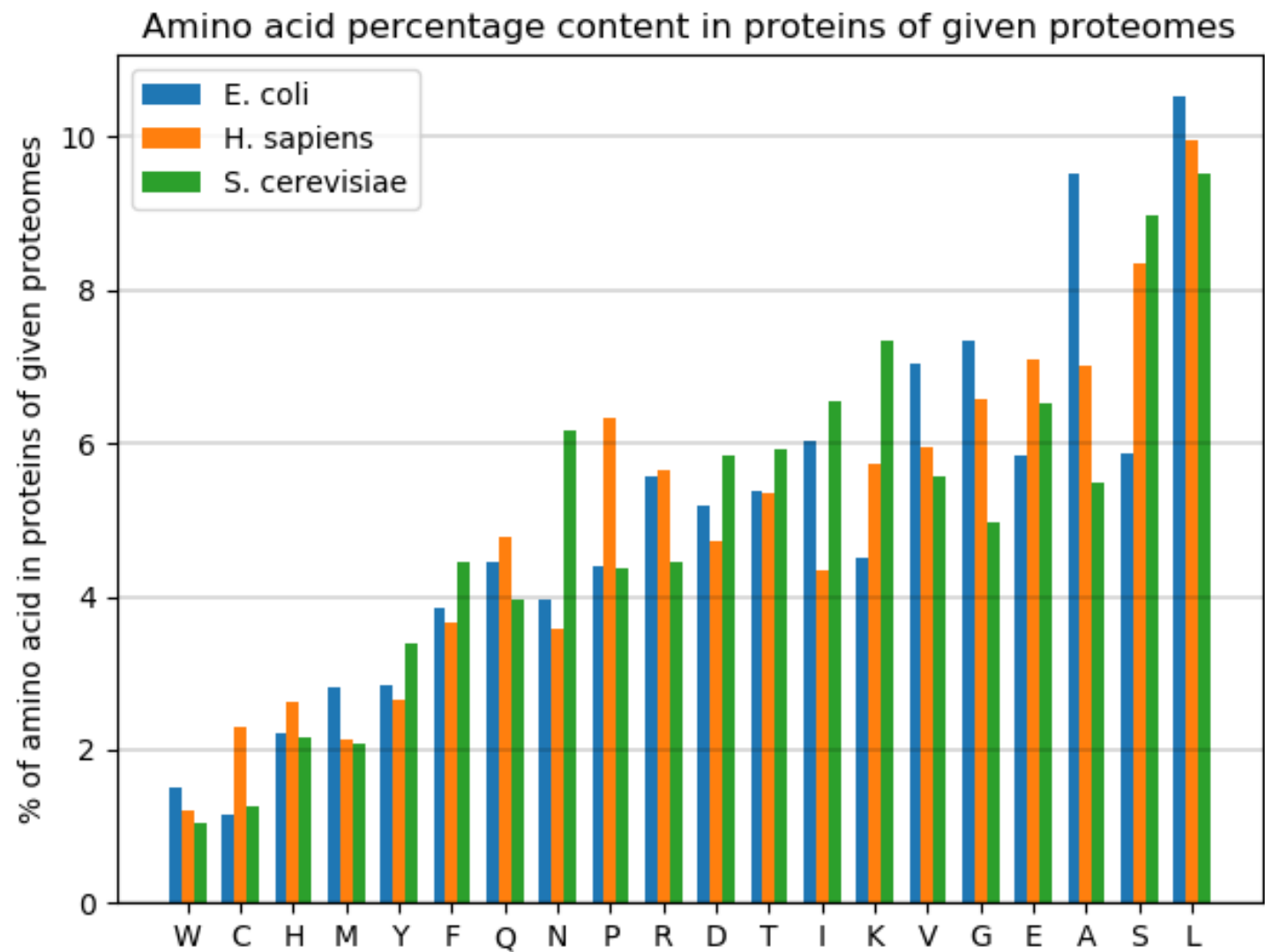


Data analysis and visualisation  
Homework I report  
Piotr Rutkowski, MIMUW

(a)



Amino acid percentage content in proteins of given proteomes									
amino acid	E. coli	C. elegans	H. sapiens	D. melanogaster	M. musculus	B. subtilis	A. thaliana	S. cerevisiae	D. rerio
L	10.54%	8.63%	9.97%	7.08%	10.12%	9.66%	9.66%	9.51%	9.5%
S	5.87%	8.08%	8.34%	7.35%	8.47%	6.28%	8.85%	8.99%	8.82%
A	9.53%	6.33%	7.01%	7.48%	6.81%	7.68%	6.48%	5.49%	6.18%
E	5.83%	6.54%	7.1%	5.49%	6.86%	7.26%	6.59%	6.52%	6.88%
V	7.04%	6.23%	5.96%	7.08%	6.12%	6.75%	6.77%	5.56%	6.28%
G	7.34%	5.35%	6.58%	6.75%	6.37%	6.91%	6.57%	4.97%	5.98%
K	4.51%	6.33%	5.72%	5.1%	5.66%	7.07%	6.24%	7.34%	5.94%
T	5.39%	5.9%	5.35%	6.82%	5.43%	5.42%	5.12%	5.91%	5.71%
I	6.02%	6.2%	4.33%	5.69%	4.48%	7.37%	5.43%	6.56%	4.73%
R	5.56%	5.14%	5.64%	7.15%	5.51%	4.09%	5.29%	4.44%	5.45%
D	5.19%	5.32%	4.73%	5.36%	4.76%	5.18%	5.34%	5.84%	5.2%
P	4.4%	4.9%	6.32%	7.08%	6.06%	3.66%	4.73%	4.38%	5.38%
N	3.95%	4.87%	3.59%	4.43%	3.61%	3.95%	4.42%	6.16%	4.05%
Q	4.45%	4.08%	4.77%	3.57%	4.72%	3.84%	3.47%	3.95%	4.72%
F	3.85%	4.77%	3.65%	3.71%	3.83%	4.5%	4.35%	4.44%	3.78%
Y	2.83%	3.21%	2.66%	2.58%	2.75%	3.49%	2.91%	3.39%	2.78%
M	2.81%	2.65%	2.13%	2.32%	2.23%	2.79%	2.5%	2.09%	2.39%
H	2.21%	2.28%	2.62%	1.85%	2.64%	2.27%	2.22%	2.17%	2.7%
C	1.14%	2.07%	2.3%	1.65%	2.37%	0.79%	1.81%	1.27%	2.39%
W	1.52%	1.11%	1.21%	1.46%	1.21%	1.03%	1.26%	1.04%	1.14%



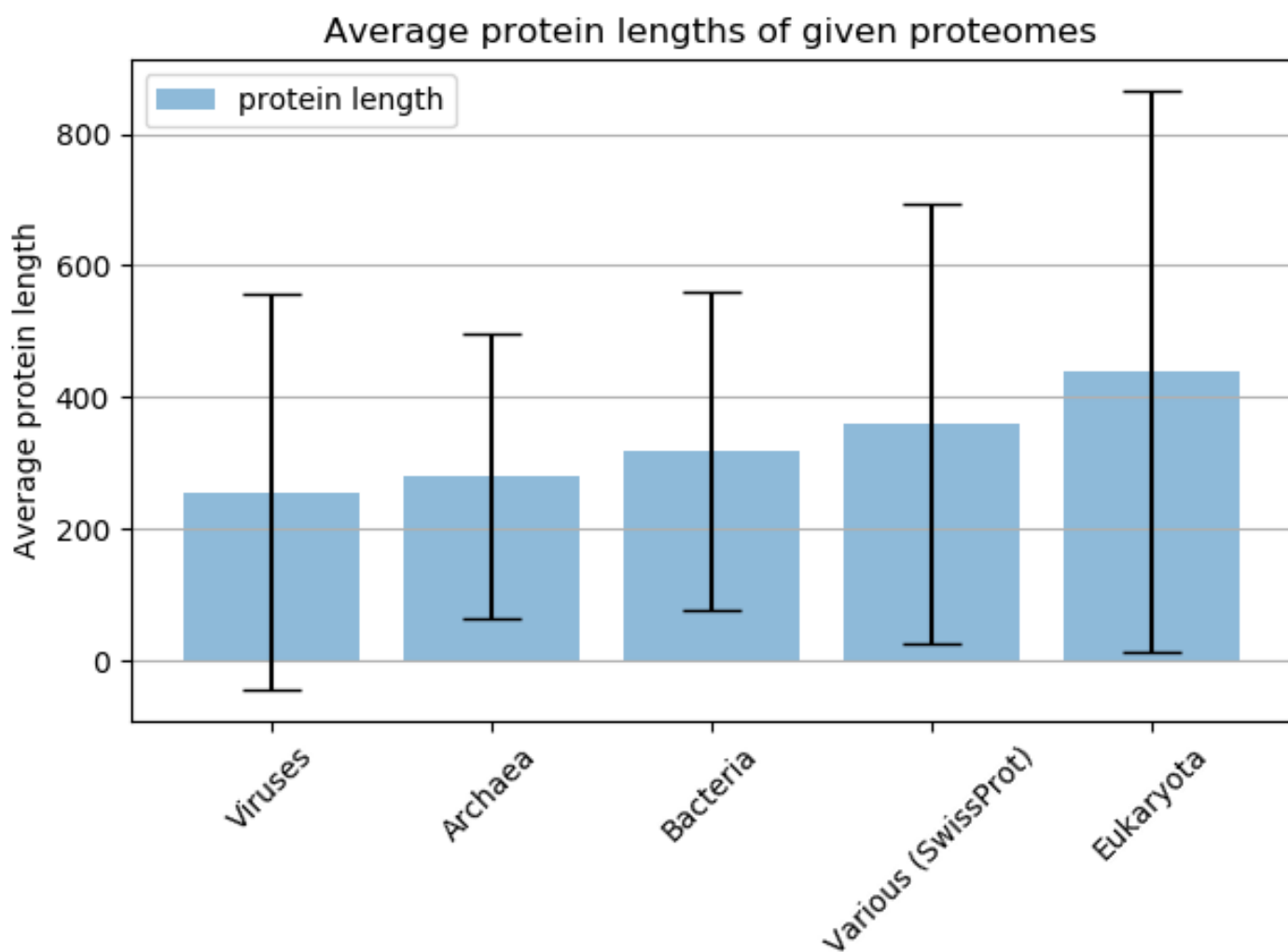
(b)

PDB average protein length:  
256.33

PDB amino acid percentage contents:

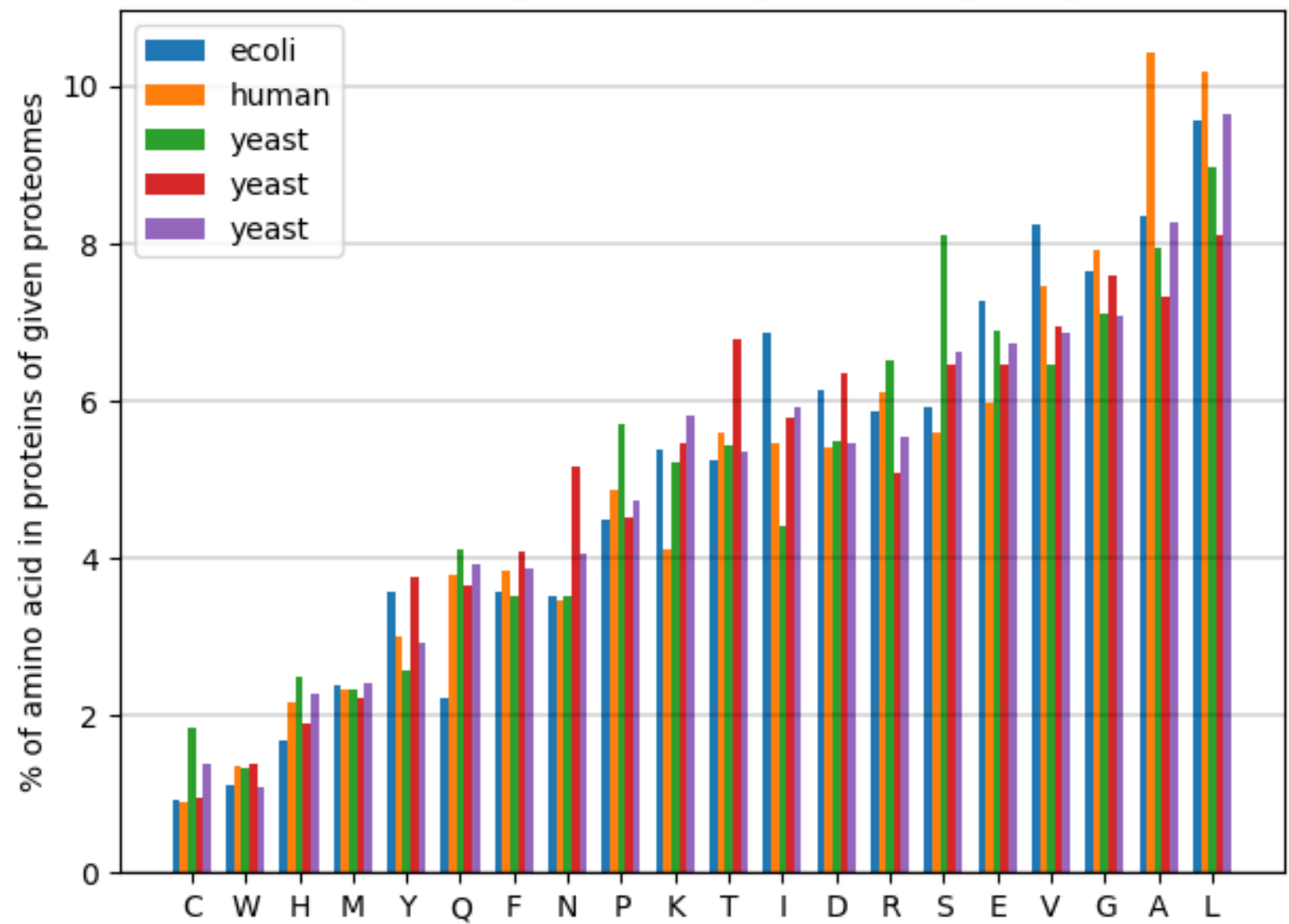
W	M	C	H	Y	Q	F	N	P	R	D	I	T	K	S	E	V	G	L	A
1.25%	2.23%	2.32%	2.5%	3.25%	3.63%	3.7%	4.03%	4.42%	5.06%	5.3%	5.33%	5.42%	5.72%	6.04%	6.27%	6.72%	8.35%	8.59%	8.63%

(c)

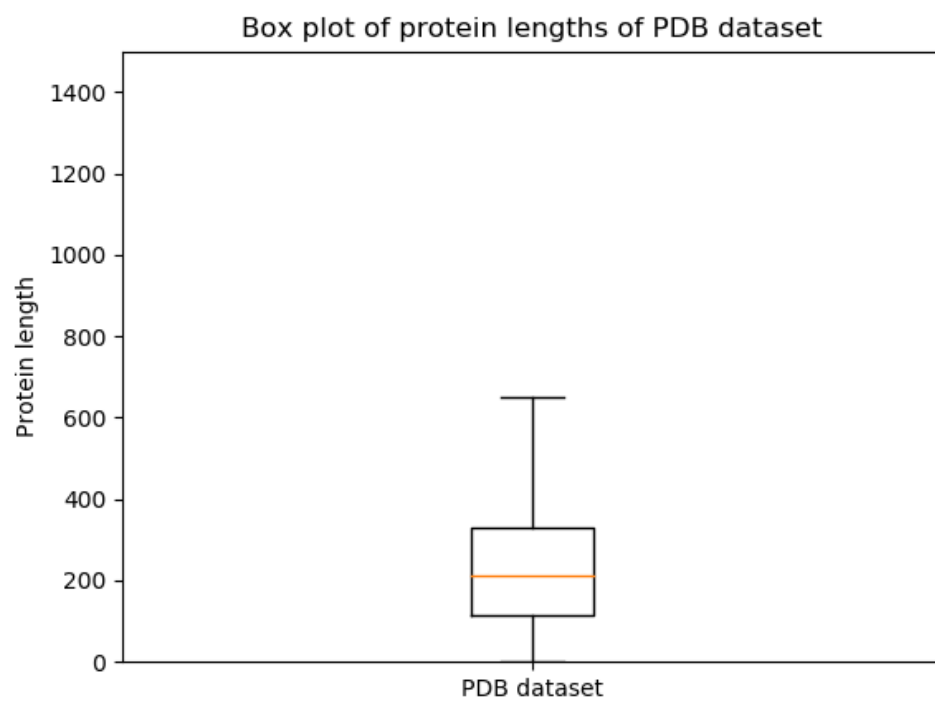
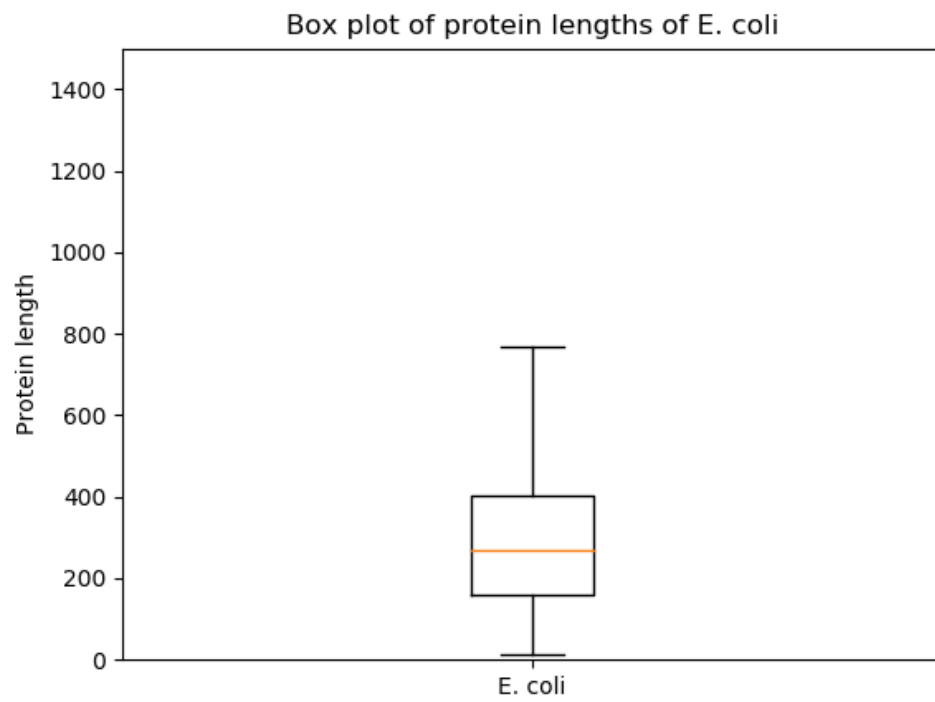


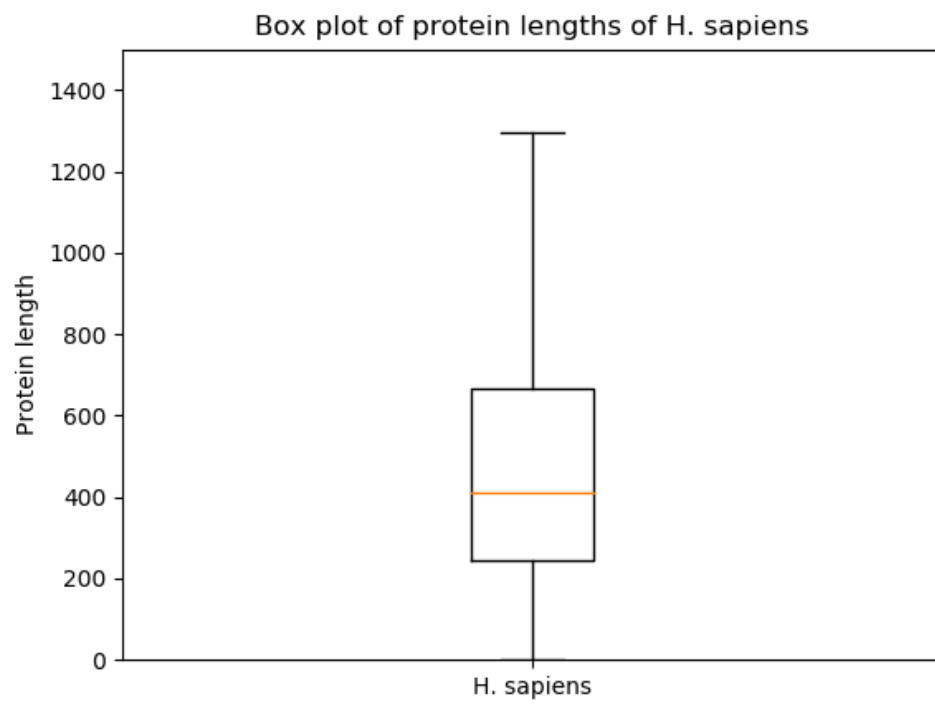
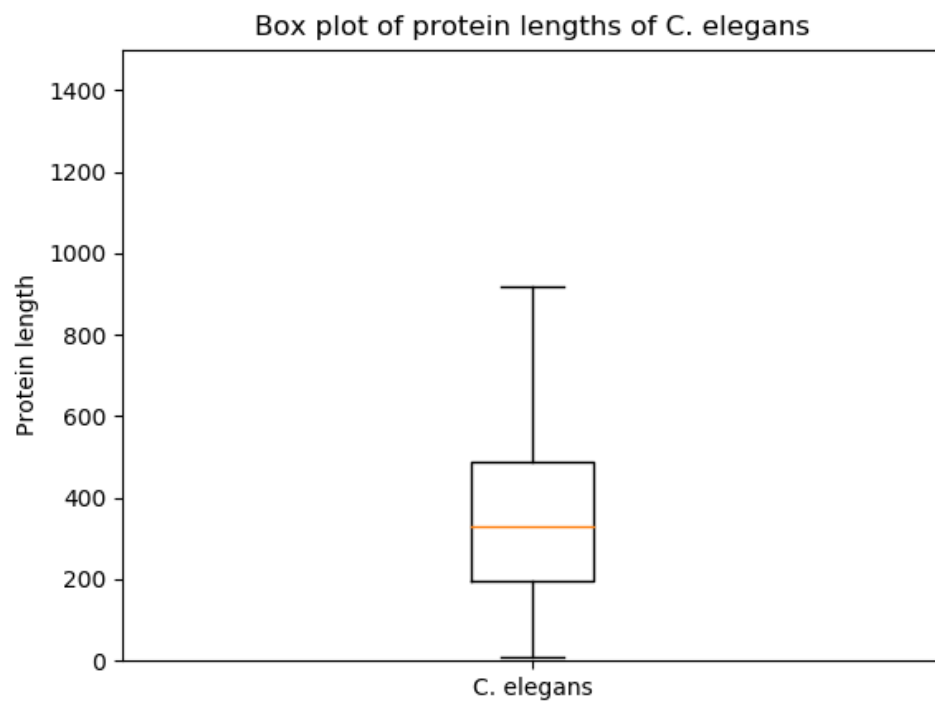
Amino acid percentage content in proteins of given proteomes					
amino acid	Archaea	Bacteria	Eukaryota	Viruses	SwissProt dataset
L	9.13%	10.16%	9.32%	8.35%	9.65%
A	8.81%	10.7%	7.67%	7.36%	8.26%
G	7.78%	8.0%	6.46%	6.51%	7.08%
V	8.09%	7.46%	6.27%	6.52%	6.86%
S	6.18%	5.74%	8.4%	6.59%	6.63%
E	7.86%	6.03%	6.4%	6.28%	6.73%
D	6.83%	5.63%	5.41%	6.26%	5.46%
I	6.17%	5.46%	5.03%	6.31%	5.92%
R	5.74%	6.18%	5.81%	5.31%	5.53%
T	5.8%	5.6%	5.55%	6.08%	5.36%
K	4.41%	4.23%	5.47%	6.2%	5.81%
P	4.33%	4.87%	5.48%	4.28%	4.74%
N	3.46%	3.33%	4.12%	5.35%	4.06%
F	3.67%	3.77%	3.83%	3.95%	3.87%
Q	2.48%	3.51%	4.15%	3.65%	3.93%
Y	3.22%	2.78%	2.82%	3.79%	2.92%
M	2.18%	2.26%	2.28%	2.43%	2.41%
H	1.86%	2.09%	2.46%	2.04%	2.28%
C	0.93%	0.88%	1.78%	1.36%	1.38%
W	1.05%	1.32%	1.27%	1.37%	1.1%

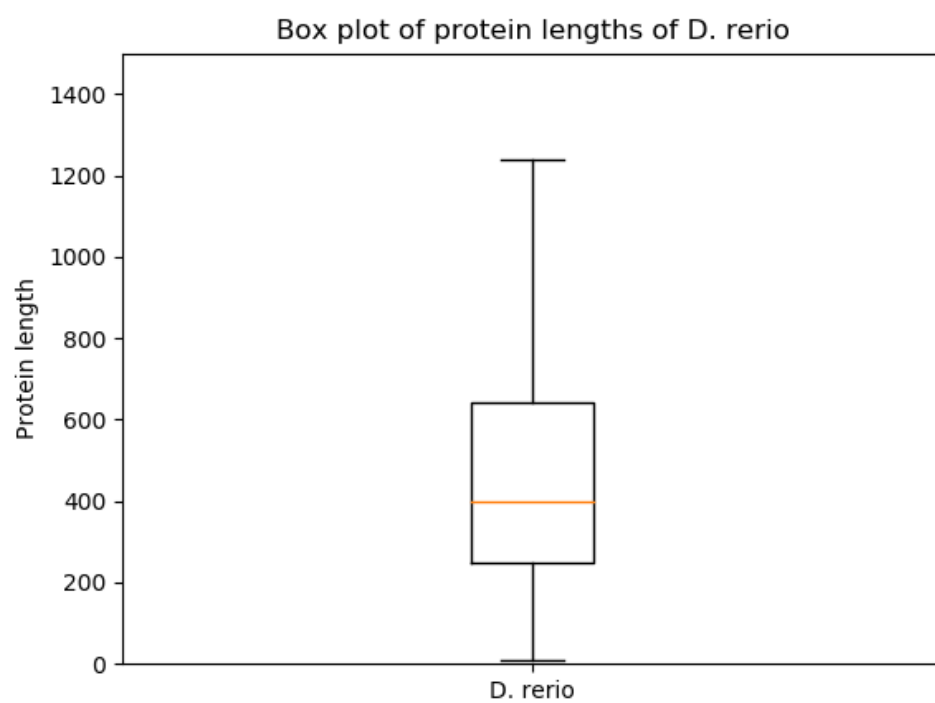
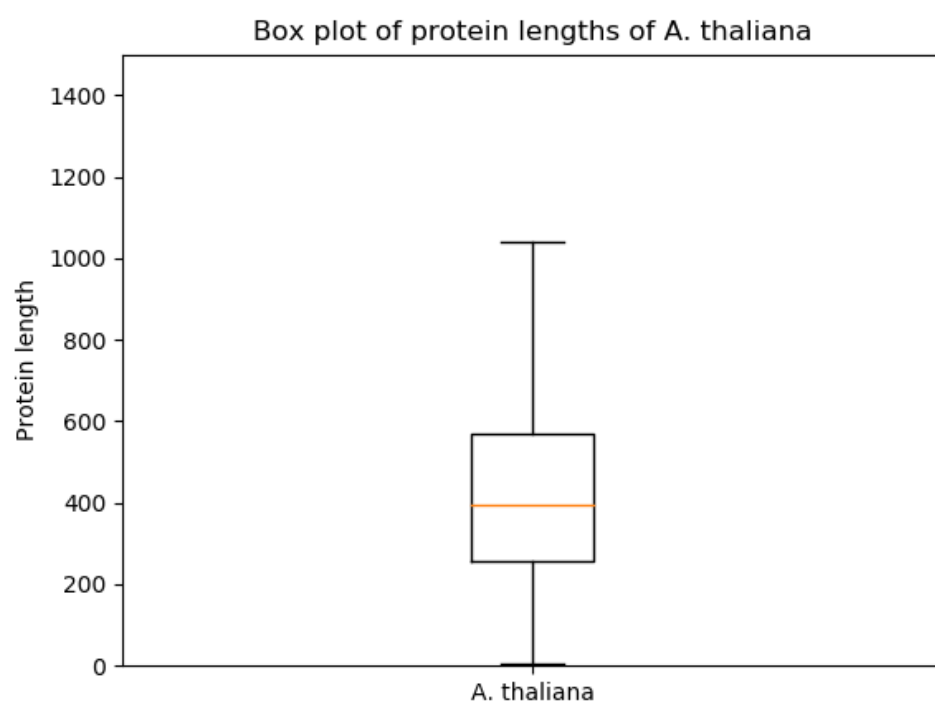
Amino acid percentage content in proteins of given proteomes



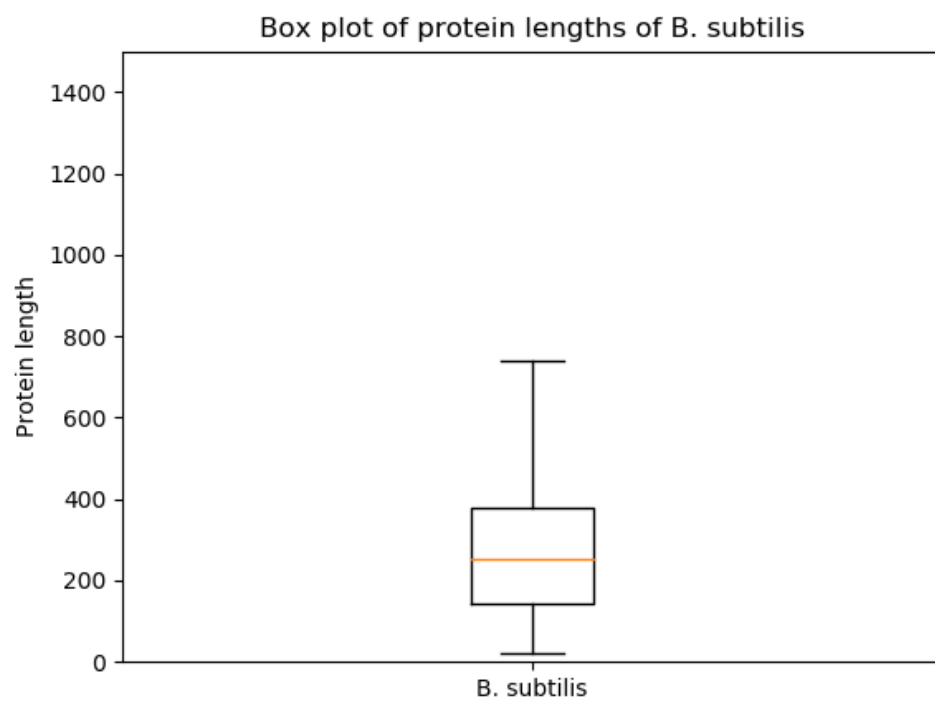
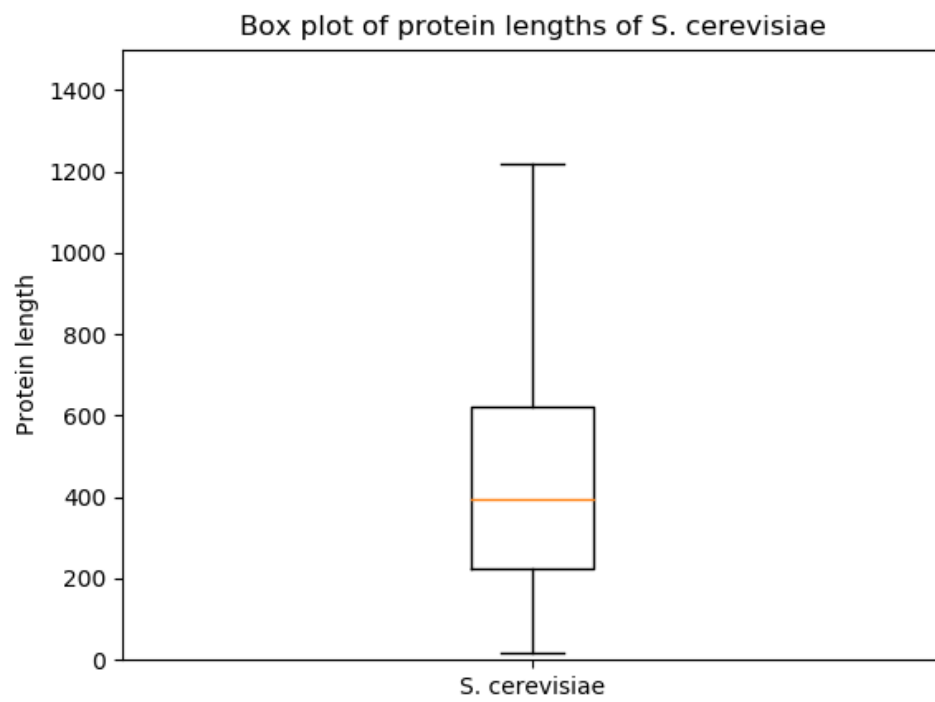
(d)

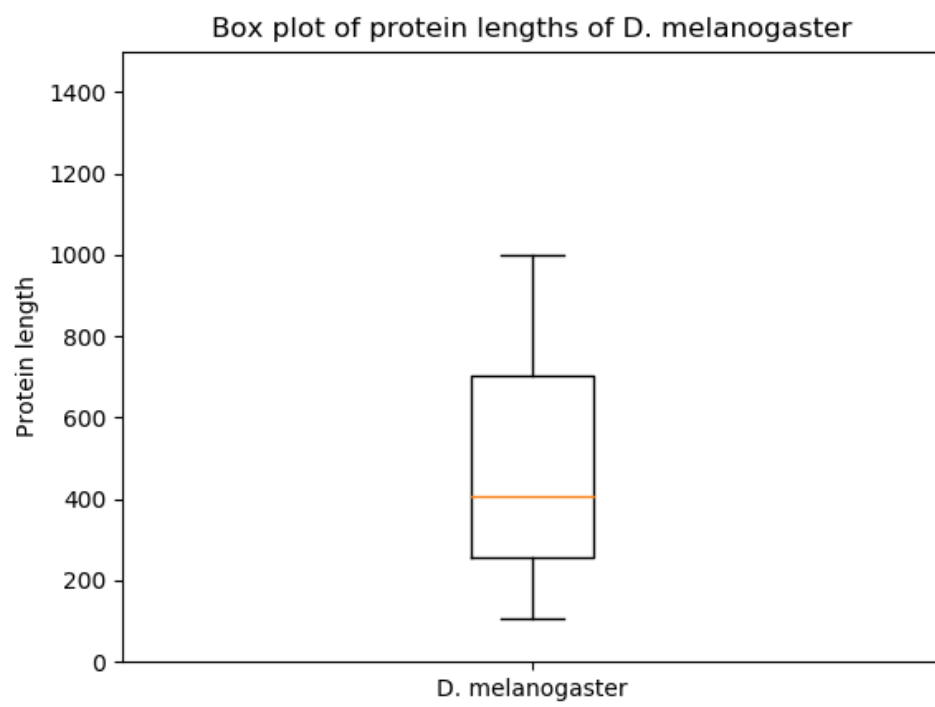
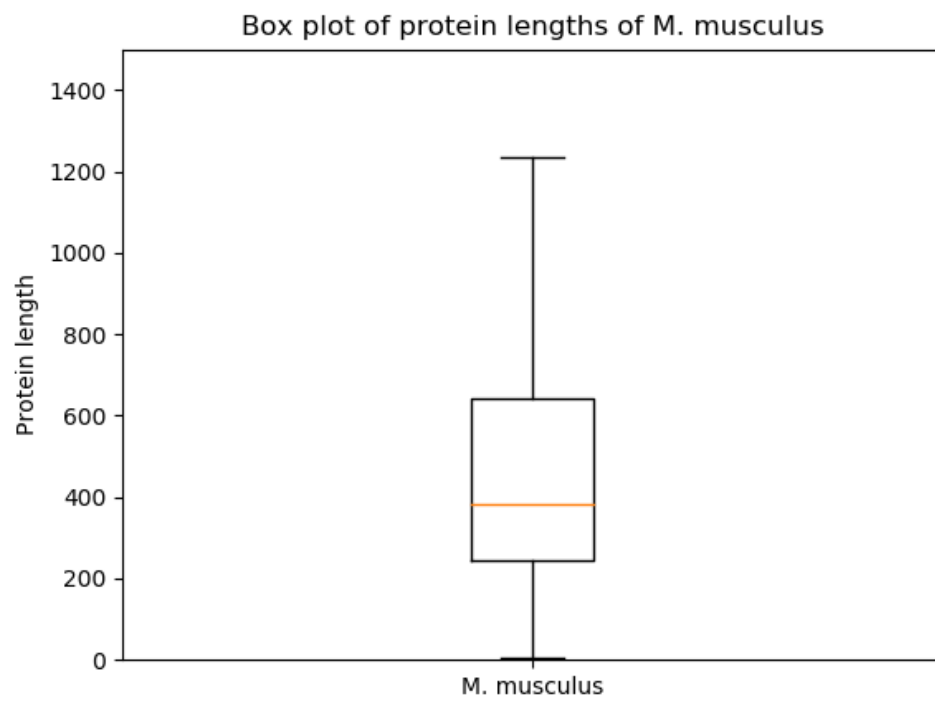












## Median vs mean

These are both values that numerically represent a set of numbers. Mean tells us what is the average value of our set, whereas median shows us where our data is split in half (50th percentile). Which value should be used is hugely dependent on the data and problem we have. The upside of mean is that it's easier to implement. But it's necessary to note that one is NOT better than the other one. If we want to achieve information on how data is split, we should use median. On the other hand, unless we want to know how data is split, we should probably use mean.

(e)

Most common amino acid at proteins' N-Terminus for given proteome	
proteome	most common amino acid at N-Terminus
E. coli	M
C. elegans	M
H. sapiens	M
D. melanogaster	M
M. musculus	M
B. subtilis	M
A. thaliana	M
S. cerevisiae	M
D. rerio	M

Methionine (M) is the most frequent amino acid at N-termini. Moreover, it's the most frequent one for all 9 organisms.