Hierarchical clustering of living organisms based on their metabolic pathways

Andrzej Małota, Tomasz Góralczyk

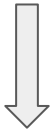
Dataset - http://bigg.ucsd.edu/models

- 108 metabolic pathways of living organisms

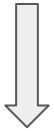
Count by kingdom:

- 77 Bacteria
- 2 Chromista
- 7 Animal
- 2 Fungi
- 6 Protozoa

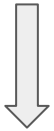
Living organism



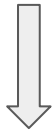
Metabolic pathway



Directed graph



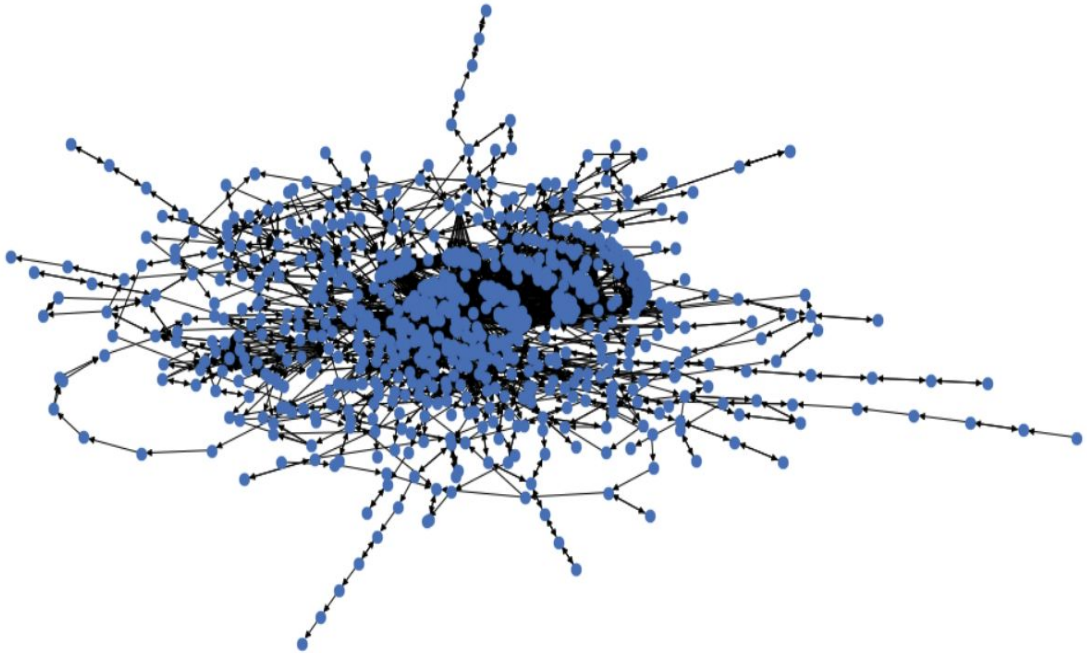
Embedded graph



Hierarchical clustering

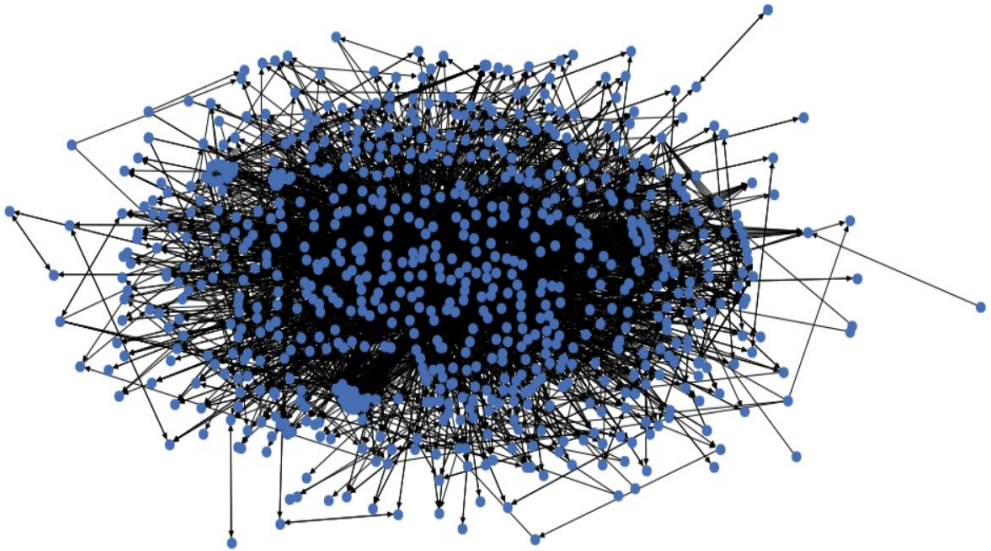
Visualisation of metabolic pathways as directed graphs

Animal - Homo sapiens - iAT_PLT_636



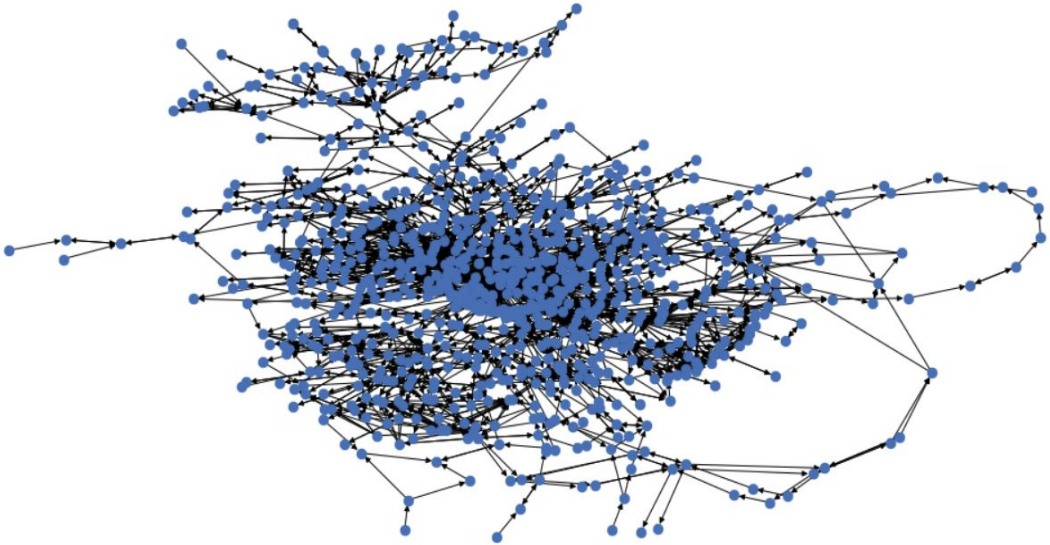
Number of nodes: 737
Number of edges: 2423
Average in degree: 3.2877
Average out degree: 3.2877
Average shortest path: 1.17
Average clustering coef: 0.145

Bacteria - Acinetobacter baumannii AYE - iCN718



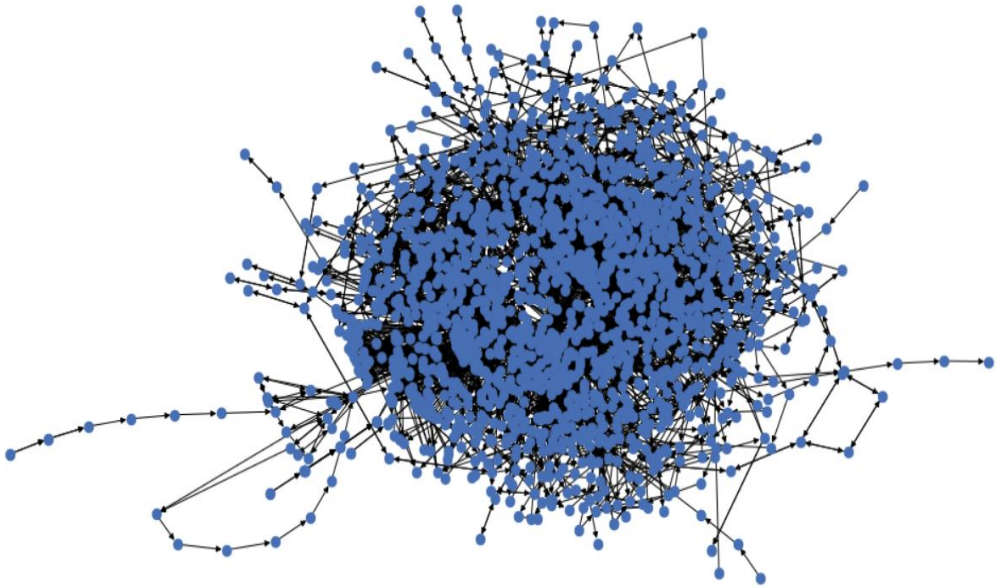
Number of nodes: 851
Number of edges: 4382
Average in degree: 5.1492
Average out degree: 5.1492
Average shortest path: 0.938
Average clustering coef: 0.202

Protozoa - Plasmodium vivax Sal-1 - iAM_Pv461



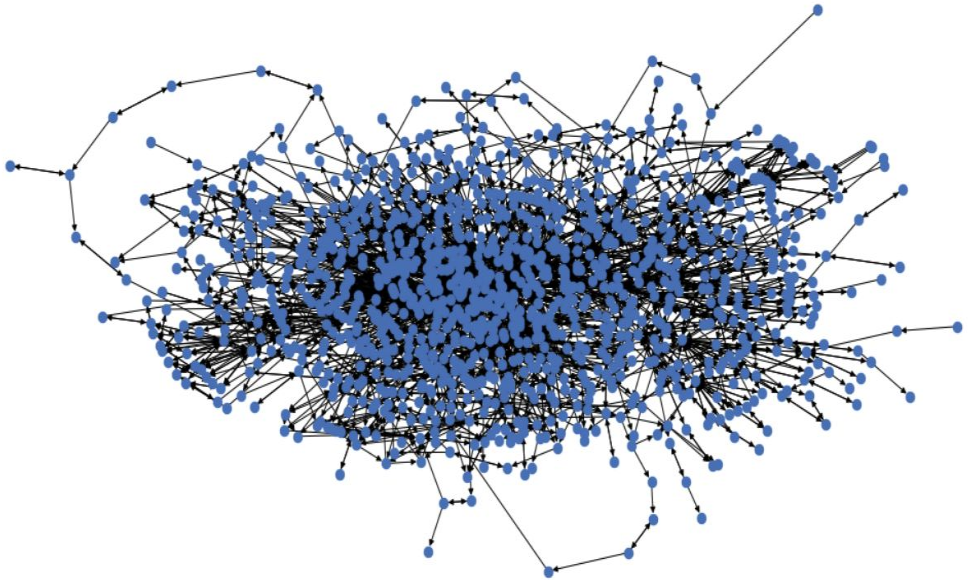
Number of nodes: 896
Number of edges: 2251
Average in degree: 2.5123
Average out degree: 2.5123
Average shortest path: 0.994
Average clustering coef: 0.139

Chromista - Chlamydomonas reinhardtii - iRC1080



Number of nodes: 1701
Number of edges: 5868
Average in degree: 3.4497
Average out degree: 3.4497
Average shortest path: 1.131
Average clustering coef: 0.168

Fungi - Saccharomyces cerevisiae S288C - iMM904



Number of nodes: 1170
Number of edges: 3207
Average in degree: 2.7410
Average out degree: 2.7410
Average shortest path: 1.121
Average clustering coef: 0.156

Graph embedding

Graph embedding

Calculate average, standard deviation, kurtosis and skewness for:

- vertex in degree
- vertex out degree
- nodes clustering coefficient
- nodes eccentricity

We also used avg. shortest path computed for each weakly component but it didn’t change the results significantly.

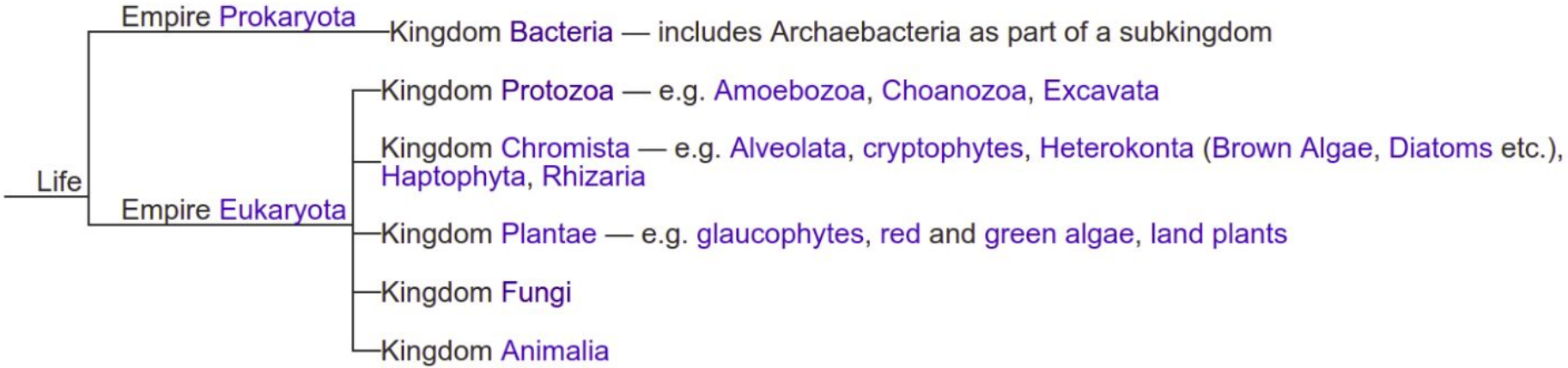
We wanted to use graph efficiency as well but networkx does not support it for directed graphs.

Features - embedded graphs

Model	avg vertex in degree	std vertex in degree	kurtosis vertex in degree	skewness vertex in degree	avg vertex out degree	std vertex out degree	kurtosis vertex out degree	skewness vertex out degree	avg clustering coefficient	std clustering coefficient	kurtosis clustering coefficient	skewness clustering coefficient
e_coli_core	6.305556	5.264237	1.401481	1.266130	6.305556	7.469640	7.532070	2.532072	0.316940	0.277065	1.014157	1.212989
iAB_RBC_283	4.388889	8.003157	96.006484	9.214740	4.388889	9.727467	125.736759	10.299814	0.135271	0.153330	3.837053	1.609463
iAF1260	4.503597	15.069085	355.906657	17.749897	4.503597	20.824420	616.767008	22.475074	0.167974	0.161343	0.467154	0.800437
iAF1260b	4.513189	15.118689	354.246729	17.716446	4.513189	20.855686	616.598394	22.468878	0.168321	0.161488	0.451520	0.795348
iAF692	4.977707	11.656010	136.549762	10.848784	4.977707	16.220205	155.000438	11.435460	0.224562	0.196762	0.368874	0.749343
iAF987	5.038774	14.345075	247.774489	14.772256	5.038774	20.668174	299.799335	15.886622	0.218038	0.179605	0.466872	0.731059
iAM_Pb448	4.397336	8.581512	179.929385	11.968657	4.397336	12.740636	277.246898	15.190135	0.139201	0.181579	5.109968	1.968645
iAM_Pc455	4.401105	8.652876	182.155984	12.085808	4.401105	12.792237	276.868382	15.193302	0.139872	0.180975	5.131952	1.963379
iAM_Pf480	4.413451	8.652928	181.695486	12.062359	4.413451	12.859943	280.351991	15.275475	0.139937	0.180297	5.204216	1.968552
iAM_Pk459	4.399118	8.629423	181.416906	12.033925	4.399118	12.806056	278.400657	15.229815	0.140272	0.181153	5.075697	1.953246
iAM_Pv461	4.393605	8.644930	182.469902	12.094873	4.393605	12.784297	277.010494	15.194741	0.139563	0.180895	5.142656	1.965763
iAPECO1_1312	4.463147	16.019030	398.014750	18.828034	4.463147	22.228411	672.589694	23.581599	0.167795	0.168157	0.364351	0.821879
iAT_PLT_636	4.940379	9.221262	144.460367	10.943656	4.940379	11.789216	269.843903	14.550248	0.145109	0.149763	2.821072	1.310503
iB21_1397	4.497910	16.266467	397.254434	18.826439	4.497910	22.470676	682.741037	23.740072	0.172907	0.169846	0.552766	0.834122
iBWG_1329	4.491146	16.241951	400.393925	18.898492	4.491146	22.350031	681.197613	23.711055	0.170318	0.166887	0.338741	0.789040
ic_1306	4.465128	16.063397	398.887502	18.865697	4.465128	22.314767	678.214056	23.703474	0.168481	0.168506	0.353557	0.815868

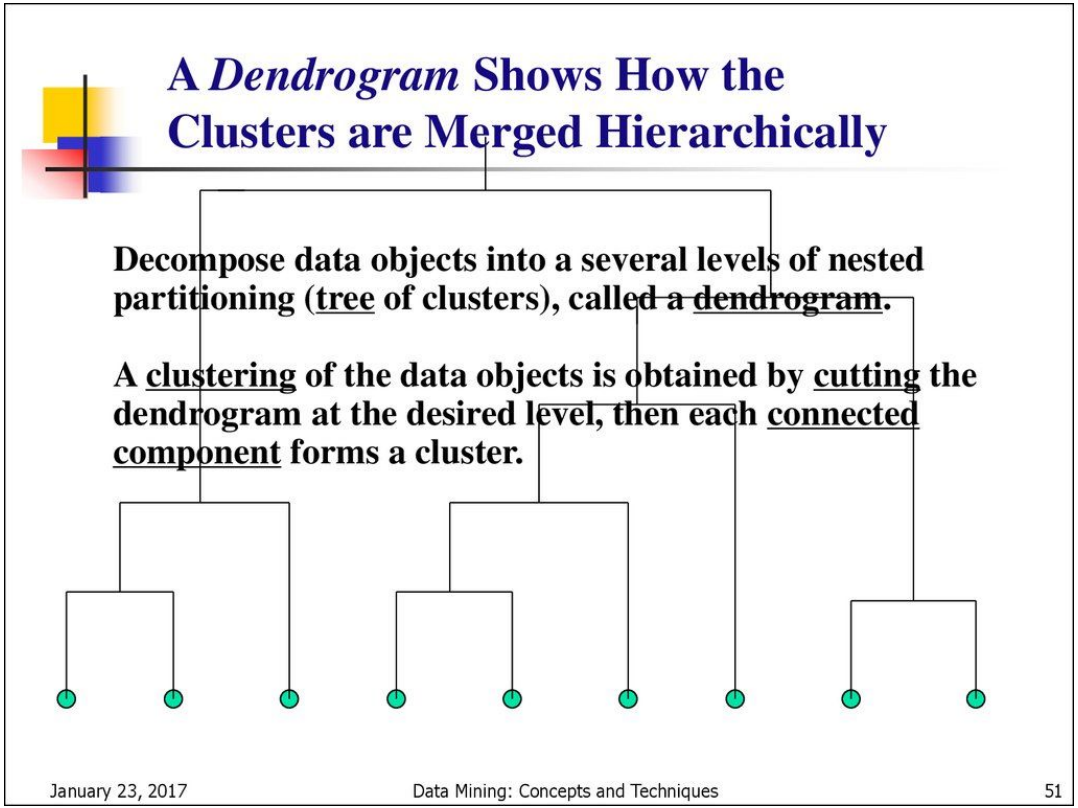
Hierarchical clustering

All organisms classified by the empire and kingdom which they belong to, from Cavalier-Smith six-kingdom model



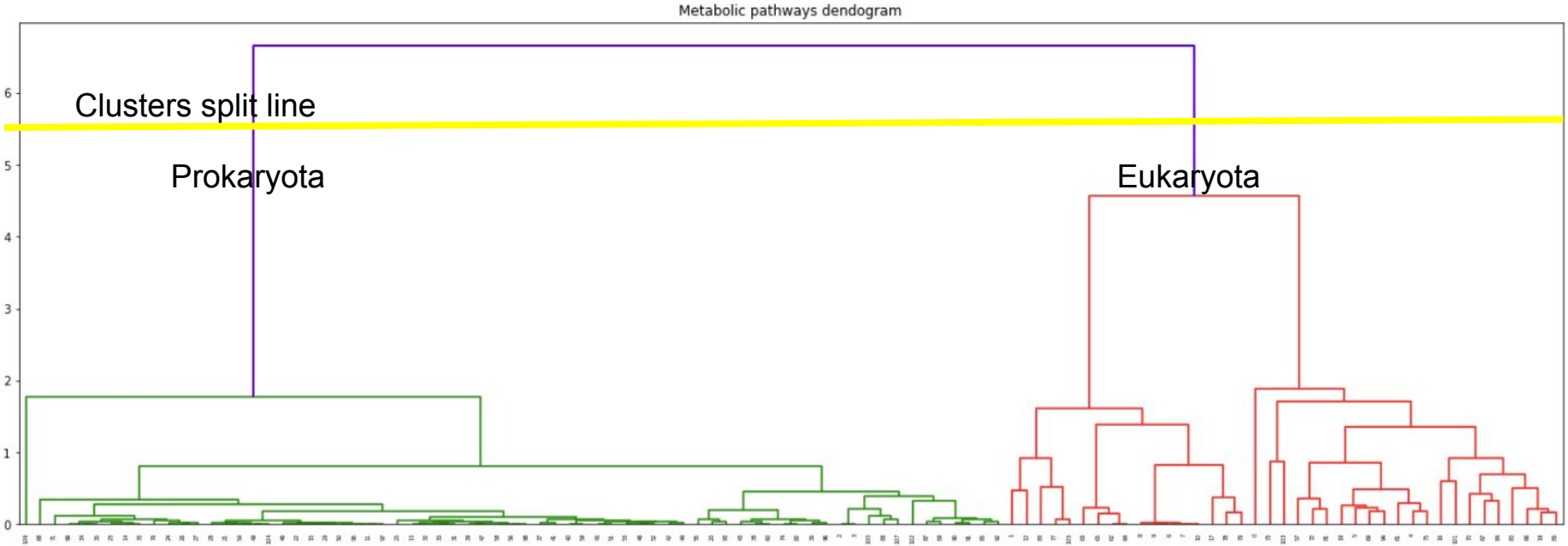
https://en.wikipedia.org/wiki/Kingdom_(biology)

Dendogram



https://en.ppt-online.org/25080

Clustering into 2 empires - dendrogram



Clustering results

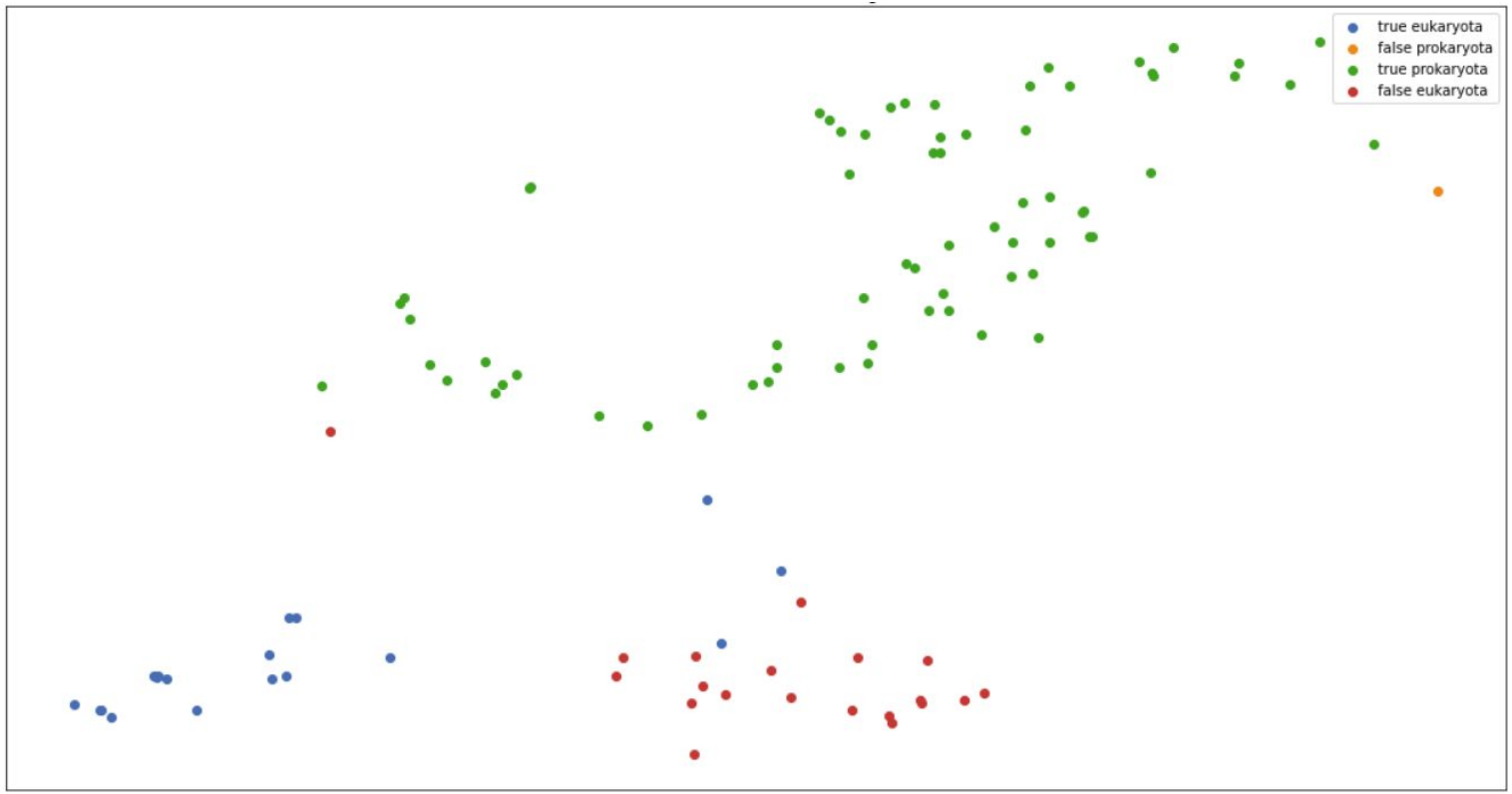
We can see that clustering was performed very successfully.

- eukaryota: 19 out of 20 eukaryota models were correctly put into cluster 0.

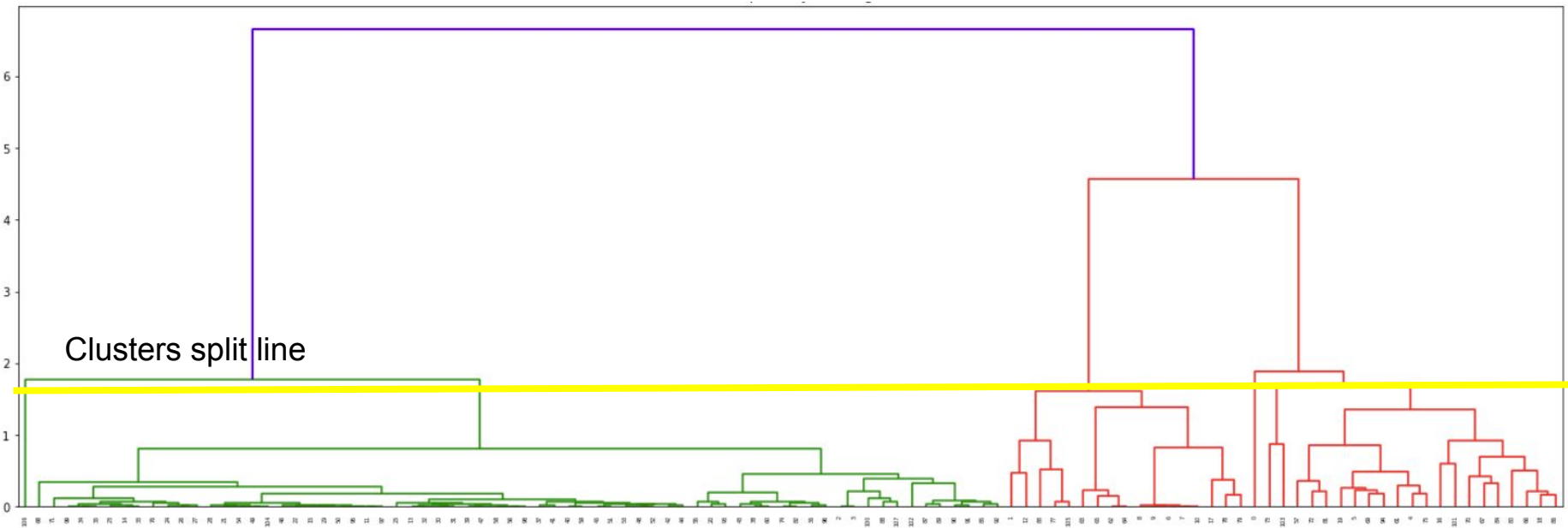
- prokaryota: 68 out of 88 prokaryota models were correctly put into cluster 1.

		models_count
empire	**cluster**	
eukaryota	**0**	19
	1	1
prokaryota	**0**	20
	1	68

Confusion matrix in 2 dimensions with t-SNE



Clustering into 6 kingdoms - dendrogram



Clustering results

- animalia: 4 out of all 7 animalia organisms were correctly put into cluster 6

- bacteria: 85 out of all 88 bacteria organisms were correctly put into cluster 1 or 2

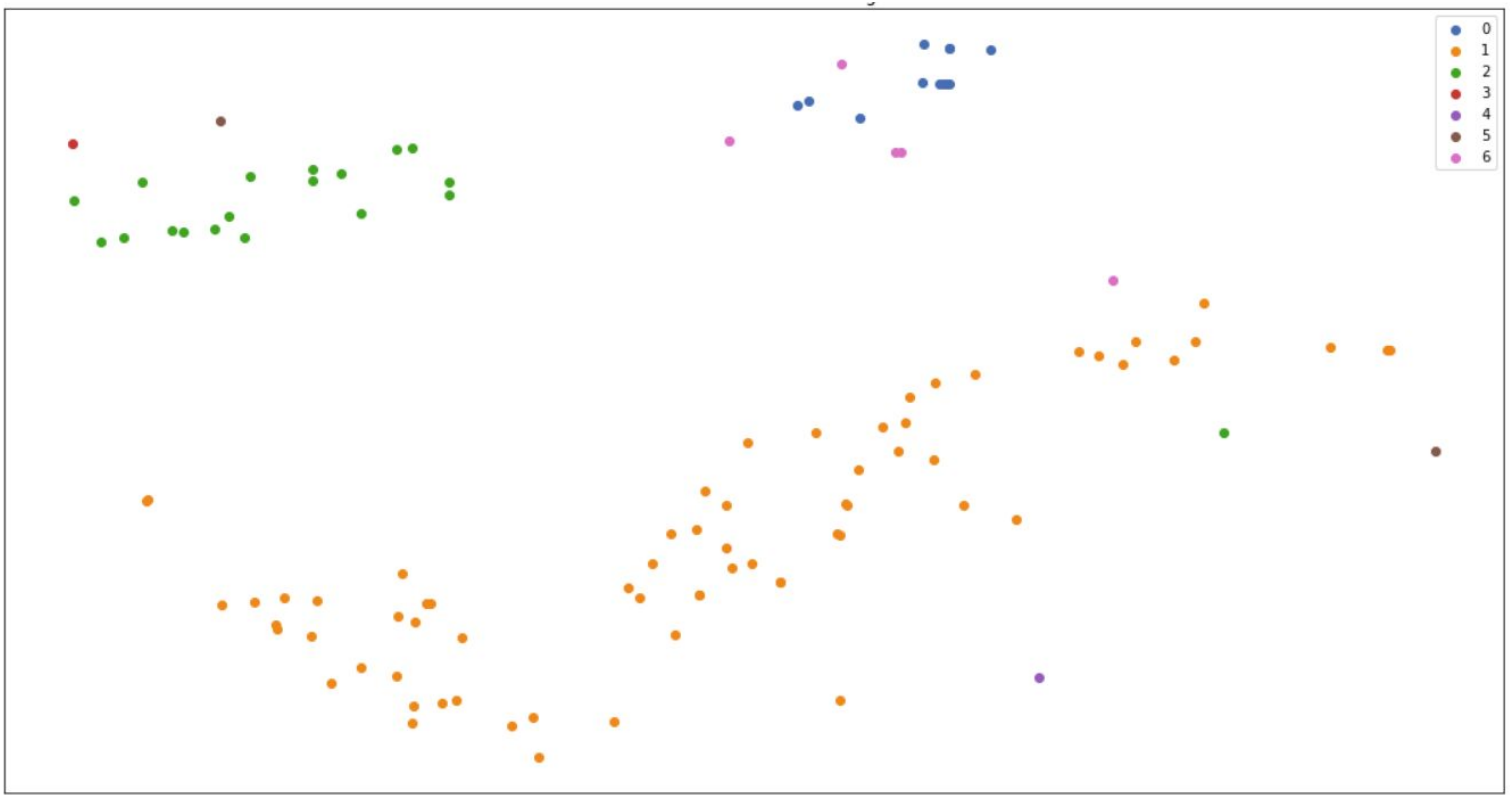
- chromista: 1 out of all 2 chromista organisms were correctly put into cluster 5

- fungi: 2 out of all 2 fungi organisms were incorrectly put into cluster 0 and should be in cluster 3

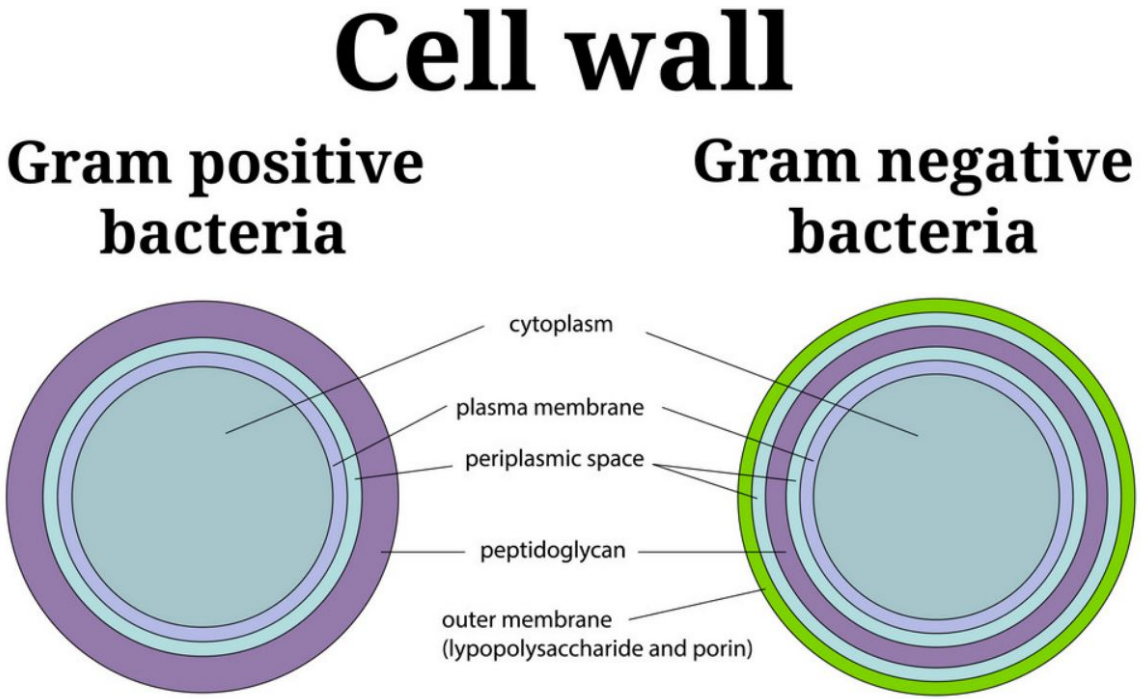
- protozoa: 9 out of all 9 protozoa organisms were correctly put into cluster 0

		models_count
kingdom	**cluster**	
animalia	**0**	1
	2	1
	4	1
	6	4
bacteria	**1**	68
	2	17
	3	1
	5	1
	6	1
chromista	**2**	1
	5	1
fungi	**0**	2
protozoa	**0**	9

Computed clusters in 2 dimension with t-SNE

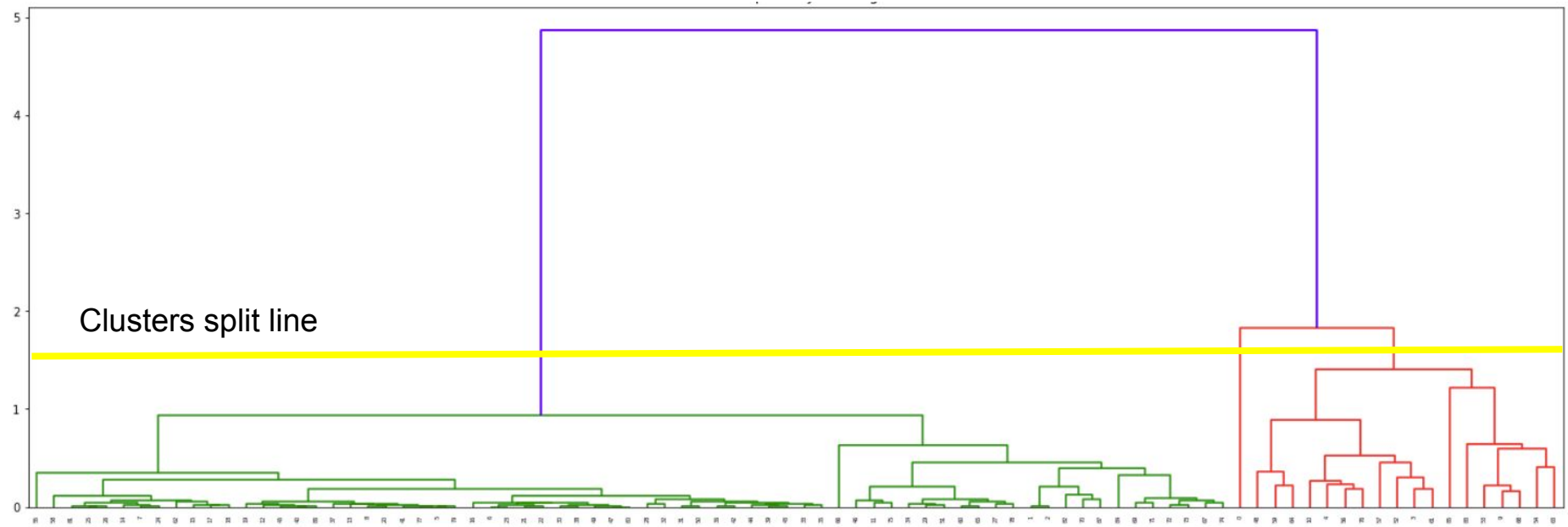


Bacteria classified by their cell wall type: gram_negative, gram_positive, gram_variable



https://www.vectorstock.com/royalty-free-vector/bacteria-cell-wall-gram-positive-and-gram-negative-vector-23107941

Clustering bacteria into 3 cell wall types - dendrogram



Clustering results

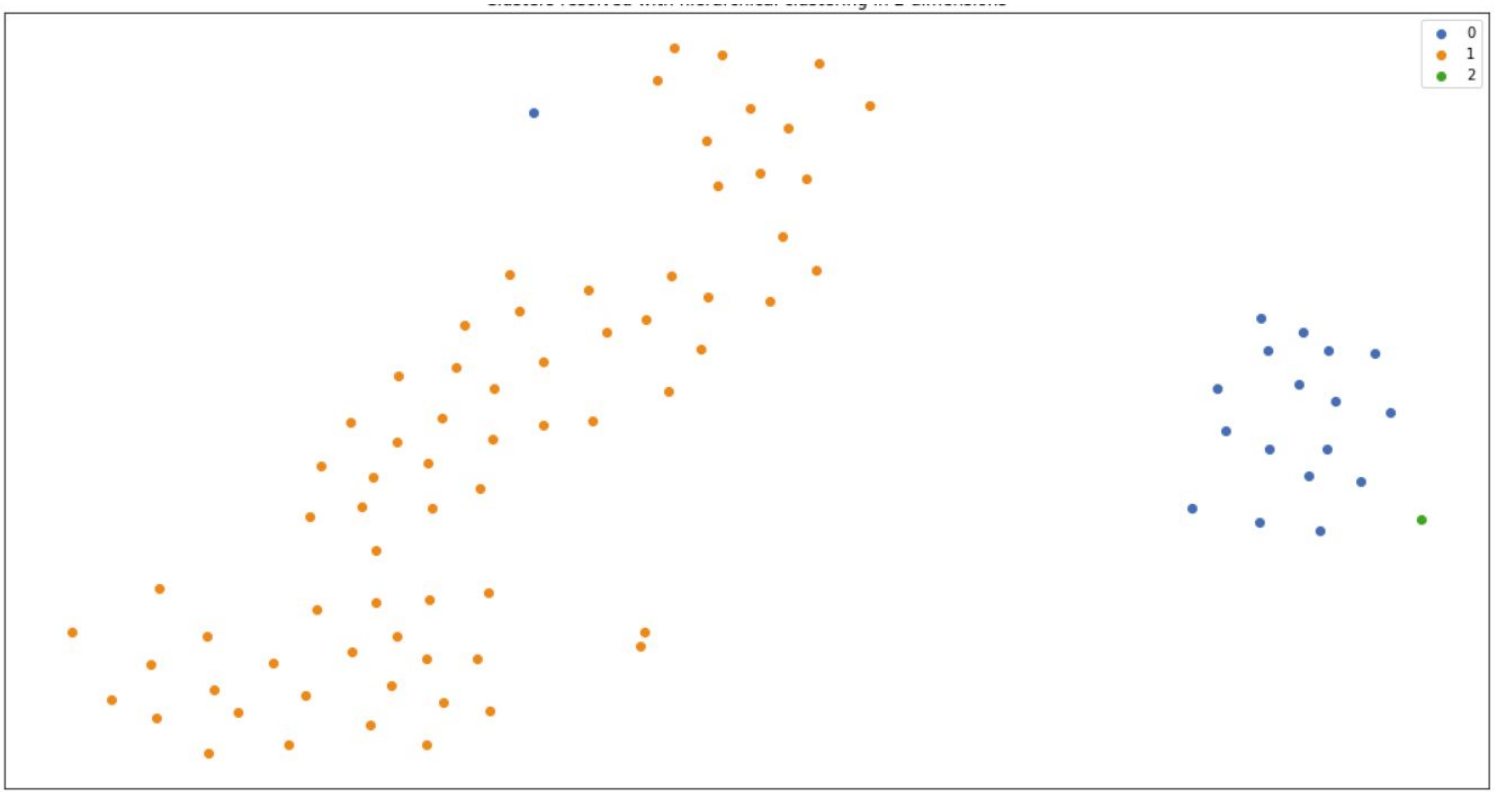
- gram_negative: 69 out of all 79 gram_negative bacteria were correctly put into cluster 1

- gram_positive: 8 out of all 8 gram_positive bacteria were correctly put into cluster 0

- gram_variable: 1 out of all 1 gram_variable bacteria were incorrectly put into cluster 0 and should be put into cluster 2

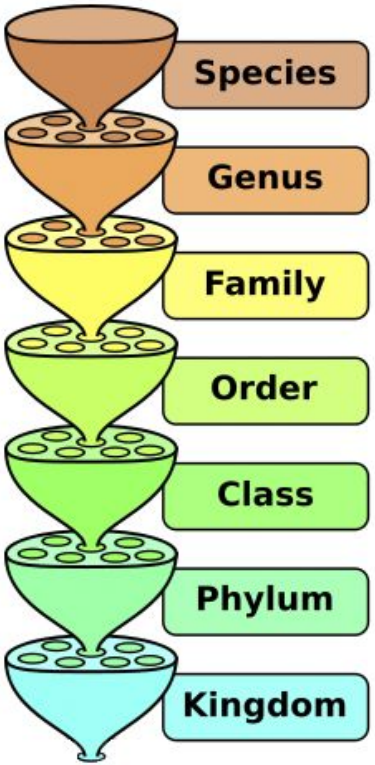
		models_count
cell_wall_type	cluster	
gram_negative	0	9
	1	69
	2	1
gram_positive	0	8
gram_variable	0	1

Computed clusters in 2 dimension with t-SNE



Comparison between calculated bacteria dendrogram and Linnaean Taxonomy of organisms

Linnaean Taxonomy



Homo sapiens
Members of the genus Homo with a hightforehead and thin skull bones.

Homo
Hominids with upright posture and large brains.

Hominids
Primates with relatively flat faces and three-dimensional vision.

Primates
Mammals with collar bones and grasping fingers.

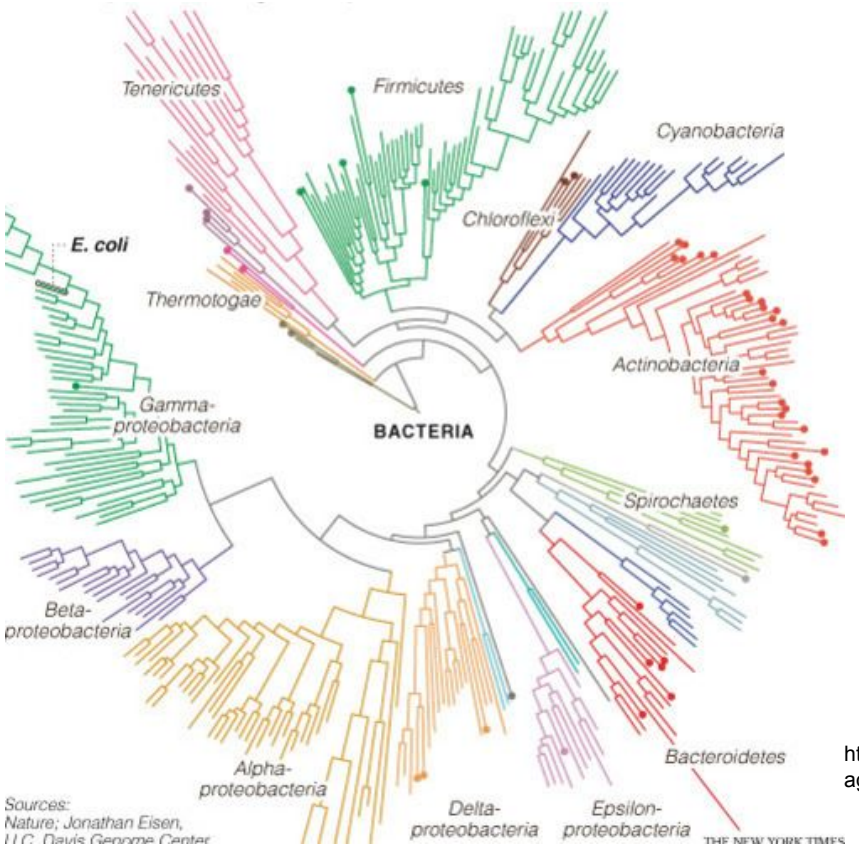
Mammals
Chordates with fur or hair and milk glands.

Chordates
Animals with a backbone.

Animals
Organisms able to move on their own.

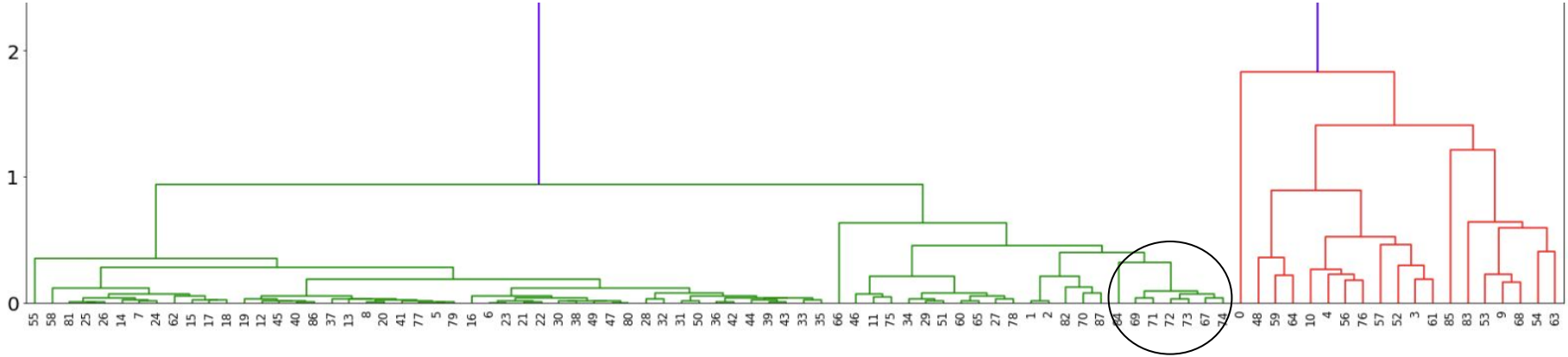
https://bhavanajagat.com/2014/06/19/spirituality-science-the-human-species/

Bacteria phylogenetic tree



https://phylogenomics.blogspot.com/2009/12/more-coverage-of-geba-phylogeny-driven.html

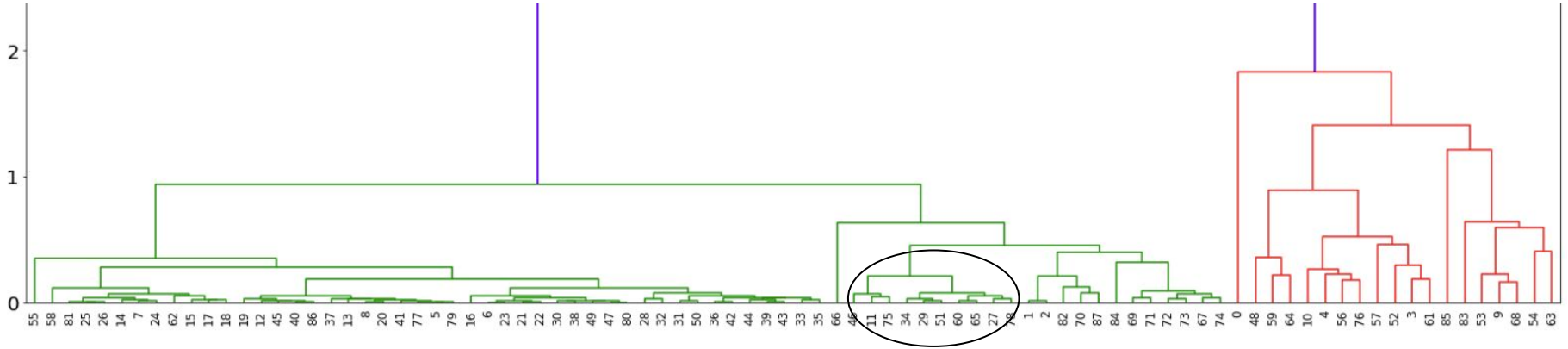
Genus - Shigella



bigg_id	gene_count	reaction_count	organism
iYS1720	1707	3357	Salmonella pan-reactome
iSbBS512_1146	1147	2591	Shigella boydii CDC 3083-94
iSDY_1059	1059	2539	Shigella dysenteriae Sd197
iSF_1195	1195	2630	Shigella flexneri 2a str. 301
iSFV_1184	1184	2621	Shigella flexneri 5 str. 8401
iS_1188	1188	2619	Shigella flexneri 2a str. 2457T
iSFxv_1172	1169	2638	Shigella flexneri 2002017

6 out of 7 bacteria from that cluster are indeed from genus shigella

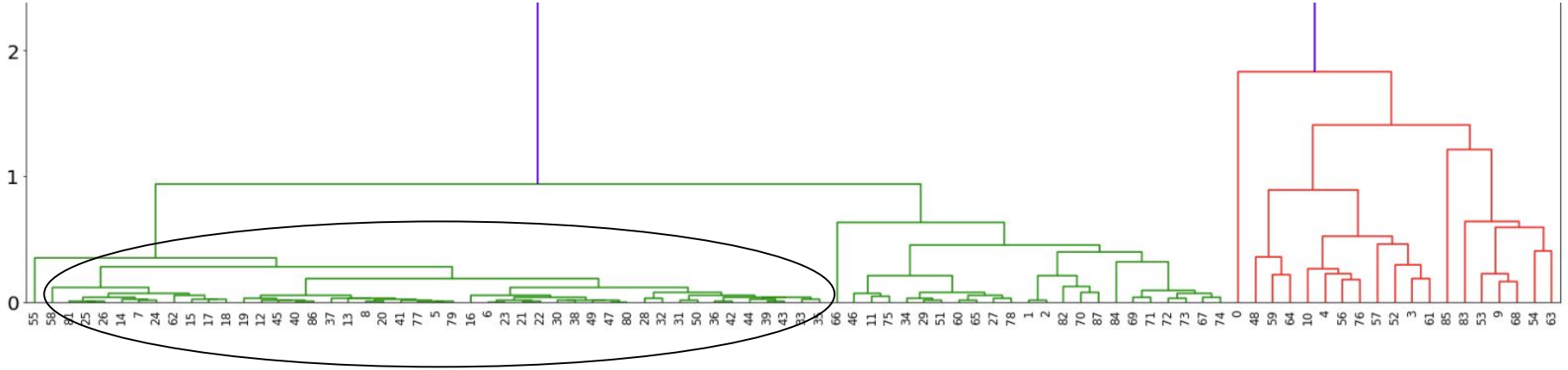
Genus - Escherichia



bigg_id	gene_count	reaction_count	organism
iECUMN_1333	1332	2740	Escherichia coli UMN026
iE2348C_1286	1287	2703	Escherichia coli O127:H6 str. E2348/69
iSSON_1240	1240	2693	Shigella sonnei Ss046
iECO103_1326	1327	2758	Escherichia coli O103:H2 str. 12009
iECH74115_1262	1262	2694	Escherichia coli O157:H7 str. EC4115
iG2583_1286	1283	2704	Escherichia coli O55:H7 str. CB9615
iLF82_1304	1302	2726	Escherichia coli LF82
iNRG857_1313	1311	2735	Escherichia coli O83:H1 str. NRG 857C
iEcE24377_1341	1341	2763	Escherichia coli O139:H28 str. E24377A
iUMNK88_1353	1353	2777	Escherichia coli UMNK88

9 out of 10 bacteria from that cluster are indeed from genus Escherichia

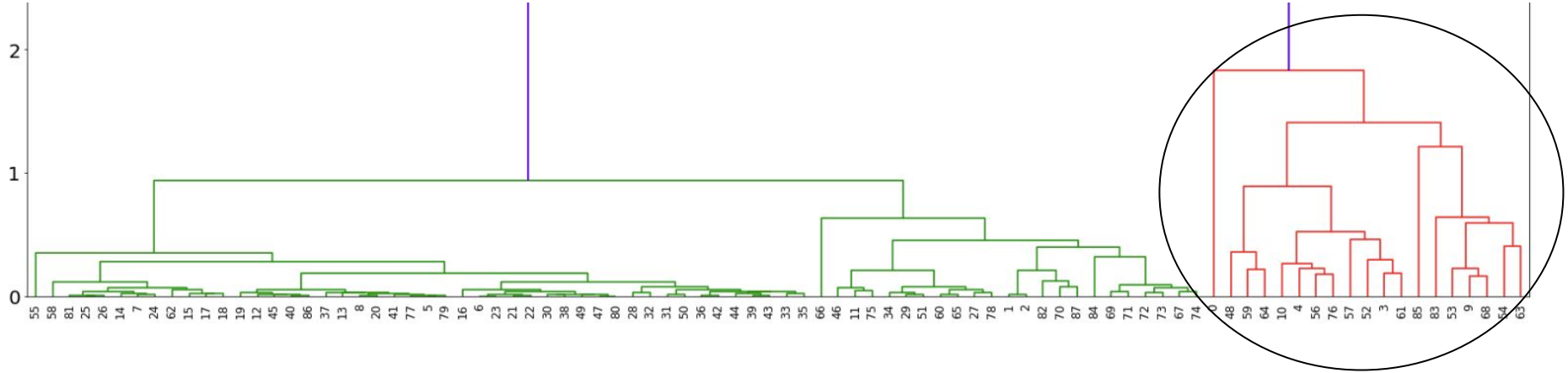
Genus - Escherichia



bigg_id	gene_count	reaction_count	organism	metabolite_count	organism_type	cell_wall_type
iJN1463	1462	2927	Pseudomonas putida KT2440	2153	Pseudomonas putida	gram_negative
iJO1366	1367	2583	Escherichia coli str. K-12 substr. MG1655	1805	Escherichia coli	gram_negative
iY75_1357	1358	2759	Escherichia coli str. K-12 substr. W3110	1953	Escherichia coli	gram_negative
iEcDH1_1363	1363	2750	Escherichia coli DH1	1949	Escherichia coli	gram_negative
iECDH1ME8569_1439	1439	2755	Escherichia coli DH1	1950	Escherichia coli	gram_negative
iEC1349_Crooks	1349	2756	Escherichia coli ATCC 8739	1946	Escherichia coli	gram_negative
iBWG_1329	1329	2741	Escherichia coli BW2952	1949	Escherichia coli	gram_negative
iECDH10B_1368	1327	2742	Escherichia coli str. K-12 substr. DH10B	1947	Escherichia coli	gram_negative
iML1515	1516	2712	Escherichia coli str. K-12 substr. MG1655	1877	Escherichia coli	gram_negative
iEC1356_Bl21DE3	1356	2740	Escherichia coli BL21(DE3)	1918	Escherichia coli	gram_negative
iEC1368_DH5a	1368	2779	Escherichia coli DH5[alpha]	1951	Escherichia coli	gram_negative
iEC1372_W3110	1372	2758	Escherichia coli str. K-12 substr. W3110	1918	Escherichia coli	gram_negative
iEC55989_1330	1330	2756	Escherichia coli 55989	1953	Escherichia coli	gram_negative
iEC042_1314	1314	2714	Escherichia coli 042	1926	Escherichia coli	gram_negative

All of the bacteria in this cluster are in phylum - Proteobacteria all the way down in hierarchy to genus - Escherichia

Phylum - can’t separate



bigg_id	gene_count	reaction_count	organism	metabolite_count	organism_type	cell_wall_type
iJN1463	1462	2927	Pseudomonas putida KT2440	2153	Pseudomonas putida	gram_negative
iJO1366	1367	2583	Escherichia coli str. K-12 substr. MG1655	1805	Escherichia coli	gram_negative
iY75_1357	1358	2759	Escherichia coli str. K-12 substr. W3110	1953	Escherichia coli	gram_negative
iEcDH1_1363	1363	2750	Escherichia coli DH1	1949	Escherichia coli	gram_negative
iECDH1ME8569_1439	1439	2755	Escherichia coli DH1	1950	Escherichia coli	gram_negative
iEC1349_Crooks	1349	2756	Escherichia coli ATCC 8739	1946	Escherichia coli	gram_negative
iBWG_1329	1329	2741	Escherichia coli BW2952	1949	Escherichia coli	gram_negative
iECDH10B_1368	1327	2742	Escherichia coli str. K-12 substr. DH10B	1947	Escherichia coli	gram_negative
iML1515	1516	2712	Escherichia coli str. K-12 substr. MG1655	1877	Escherichia coli	gram_negative
iEC1356_Bl21DE3	1356	2740	Escherichia coli BL21(DE3)	1918	Escherichia coli	gram_negative
iEC1368_DH5a	1368	2779	Escherichia coli DH5[alpha]	1951	Escherichia coli	gram_negative
iEC1372_W3110	1372	2758	Escherichia coli str. K-12 substr. W3110	1918	Escherichia coli	gram_negative
iEC55989_1330	1330	2756	Escherichia coli 55989	1953	Escherichia coli	gram_negative
iEC042_1314	1314	2714	Escherichia coli 042	1926	Escherichia coli	gram_negative

Bacteria in this cluster can’t be appropriately separated on any level up to phyla where bacterias belong to phylums: Proteobacteria, Actinobacteria, Firmicutes, Cyanobacteria,Thermotogae

Summary

In this study we managed to:

- Convert metabolic pathways into directed graphs,
- Embed resulting graphs into feature vector,
- Label each organism by its empire and kingdom and then very successfully cluster them into empires and kingdoms they belong to with the help of hierarchical clustering,
- Filter out bacteria from all organisms and label each bacteria by its cell wall type and then with great results cluster with hierarchical clustering.
- Additionally we performed the clustering with avg. shortest path feature but the results did not improve compared to our baseline

Linnaean Taxonomy comparison results

- We were able to identify 2 clusters with its members genus - Escherichia (blue),
- We were able to identify cluster with its members genus - Shigella (black),
- For the bacteria on the red branch we couldn’t find clusters with its members that belong to same group on any level in Linnaean Taxonomy (red).