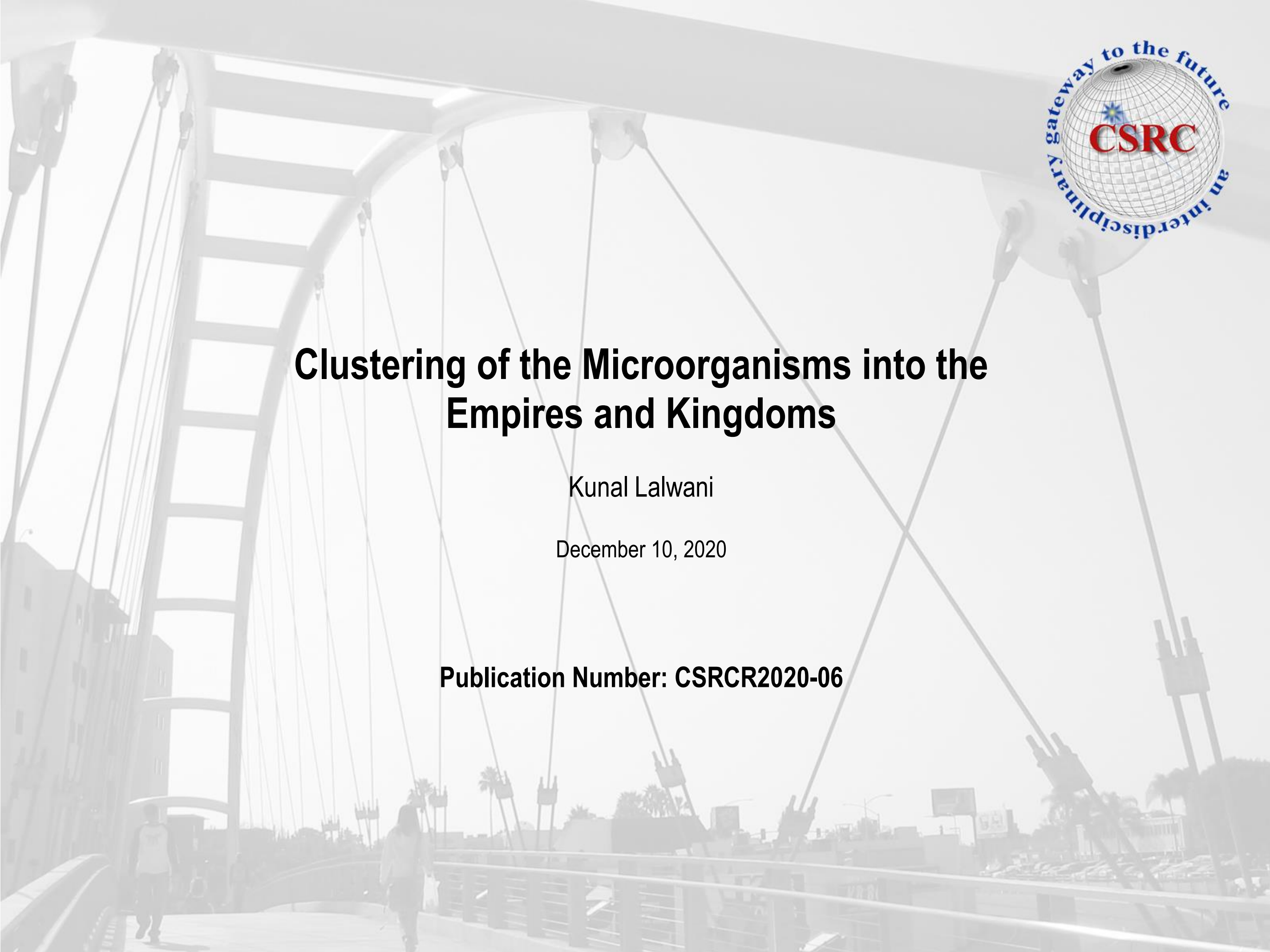


Clustering of the Microorganisms into the Empires and Kingdoms

Kunal Lalwani

December 10, 2020

Publication Number: CSRCR2020-06

SAN DIEGO STATE
UNIVERSITY

Computational Science Research Center
College of Sciences
5500 Campanile Drive
San Diego, CA 92182-1245
(619) 594-3430

Clustering of the Microorganisms into the empires and kingdoms

A Project

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

In

Computational Science

by

Kunal Lalwani

Fall 2020

Table of Contents

Abstract

The metabolic pathways is studied to help understand the functionalities inside an organism. In this study we used the metabolic pathways to group certain sets of species based on their hierarchy. The BiGG Models Platform is an extensive platform for the genomic dataset and where the genomic structures and the metabolic pathway information about 108 organisms are stores. Cameo package [16] in python contained this dataset as a library and was used for network mesh analysis. A directed graph (DAG) is constructed from this network mesh having the indegree and outdegree vertices count and also the clustering coefficient count. The DAGs were transformed to an embedding using statistical techniques which were used as feature vectors for the machine learning algorithm. The hierarchical clustering was performed as different sets of experiments and studied. The number of clusters were taken as an idea from Dendrogram- which is a hierarchical tree structure representation of the datapoints. Finally, a label to each organism by its empire and kingdom were produced and then very successfully cluster them into empires and kingdoms they belong to with the help of hierarchical clustering. There was another experiment to filter out bacteria from all organisms and label each bacterium by its cell wall type and then with great results cluster with hierarchical clustering. Additionally, the clustering with avg. shortest path feature was also performed, but the results did not improve compared to the baseline. A common phylum containing genus like Escherichia and Shigella were clearly identified in the clustering proving the validity of the clustering technique. Whereas there were some species whose phylum levels were difficult to comment from the clusters due to the rareness and less abundance data entries.

Chapter 1

Introduction

1.1 Metabolic Pathways

A metabolic pathway is a connected arrangement of synthetic responses happening inside a cell. The reactants, items, and intermediates of an enzymatic response are known as metabolites, which are changed by a grouping of compound responses catalyzed by enzymes. In many instances of a metabolic pathway, the result of one chemical goes about as the substrate for the following. In any case, side items are viewed as waste and taken out from the cell. These chemicals frequently require dietary minerals, nutrients, and different cofactors to work.

Each metabolic pathway comprises of a progression of biochemical responses that are associated by their intermediates: the results of one response are the substrates for ensuing responses, etc. Metabolic pathways are regularly considered to stream one way. Although all synthetic responses are reversible, conditions in the phone are regularly with the end goal that it is thermodynamically more positive for motion to continue one way of a response.

1.2 Open-Source Databases

With the growing revolve around sequencing advancement, natural science has in like manner gotten dynamically subject to programming designing to gainfully store and access data, provoking the creation of databases to house this information. An extraordinary piece of the information is moreover made available to examiners through open-access or open databases, proposed to empower analysts to get to likewise disseminate their own data. Anyway, just creation this data open is not adequate. People who wish to use it ought to in like manner have the data and ability to recognize and download the data they search for. This concentrations to a creating necessity for mechanical assemblies and shows to help the people who may use this bounty of open access data do so successfully

1.3 The BiGG Models Platform

Genome-scale metabolic models' unit mathematically structured data bases that is throughout a really difficult position to be accustomed predict metabolic pathway usage and growth phenotypes. Moreover, they're going to induce, and check hypotheses once integrated with experimental data. to maximize the worth of those models, centralized repositories of high-quality models got to be established, models got to adhere to established standards and model components got to be connected to relevant databases. Tools for model visual image heaps of enhance their utility.

BiGG Models gives an exhaustive application programming interface for getting to BiGG Models with showing and examination devices. As a resource for very curated, standardized and open models of assimilation, BiGG Models empowers various structures science studies and sponsorship data-based examination of various exploratory information.

1.4 Dataset and Packages

BiGG Models can be accessed using a simple web API. Some instructions were followed as present on http://bigg.ucsd.edu/data_access to download the files in xml format. The model was in the form of a COBRA model which could be easily parsed by the COBRA package in python to do anaylsis. Cameo package [16] in python was used which is an extension of cobrapy but with extra features like loading models from different models- in my case from .xml format, direct access to BiGG models, using the optlang solver interface to optimize the problem which is based on sympy package in python.

1.5 Project Aim

A clustering method to predict the empire and kingdom of the organisms. The hierarchical clustering in an unsupervised Machine Learning technique, an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. This clusters were used to identify the kingdoms and empires of the organism and able to differentiate when comparing with the Linnean Taxonomy.

1.5 Linnaean Taxonomy

The Linnaean arrangement of order comprises of a progression of groupings, called taxa (singular, taxon). Taxa range from the realm to the species. The kingdom is the biggest and most comprehensive gathering. It comprises of creatures that share only a couple essential similarities. The species is the simplest and most selective gathering. It comprises of creatures that are comparable enough to deliver ripe posterity together. Firmly related species are gathered in a class

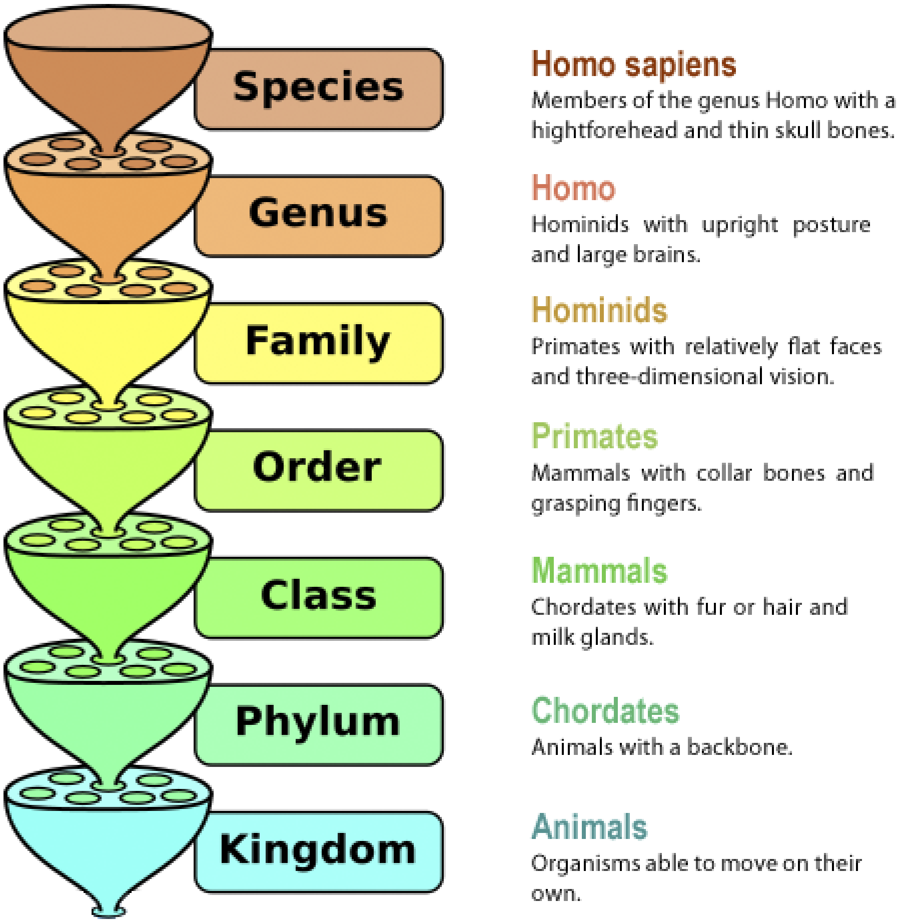

Species
Homo sapiens
Members of the genus Homo with a hightforehead and thin skull bones.
Genus
Homo
Hominids with upright posture and large brains.
Family
Hominids
Primates with relatively flat faces and three-dimensional vision.
Order
Primates
Mammals with collar bones and grasping fingers.
Class
Mammals
Chordates with fur or hair and milk glands.
Phylum
Chordates
Animals with a backbone.
Kingdom
Animals
Organisms able to move on their own.

Figure 1: spirituality science – the human species: Linnaeus divided organisms into two kingdoms – 1. Animalia(animals), and 2. Plantae(plants). The kingdoms are divided into the following categories: phylum or division, class, order, family, genus, and species us

1.5 Gram positive vs Gram Negative Bacteria

Most microorganisms are requested into two general groupings: Gram positive and Gram negative. These classes rely upon their cell divider association and reaction to the Gram stain test. The Gram recoloring strategy, made by Hans Christian Gram, perceives minute living beings reliant on the reaction of their telephone dividers to explicit tones and artificial materials.

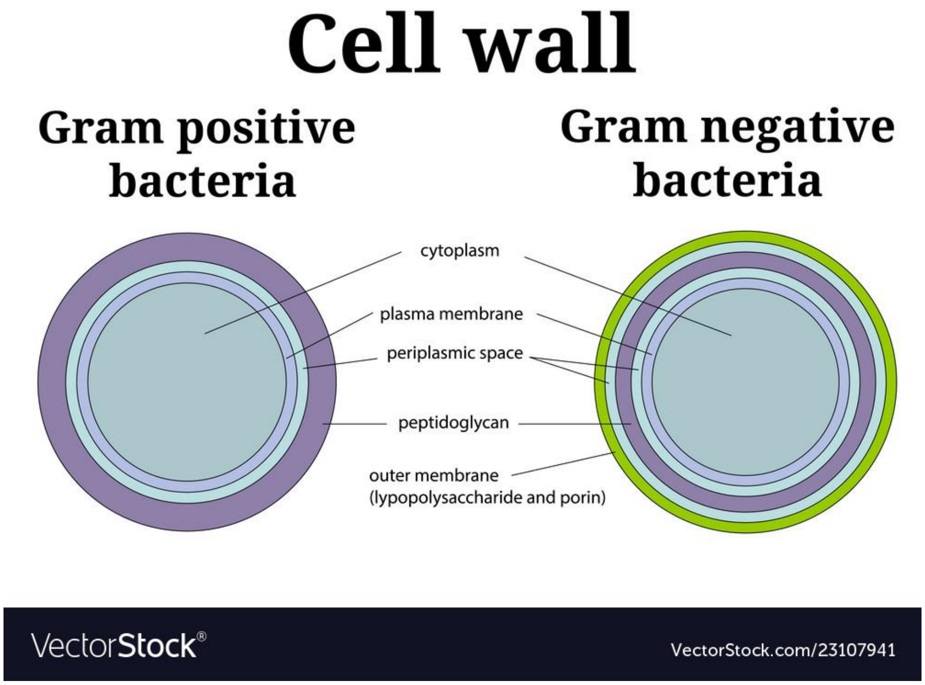


Figure 2: Bacteria - classified by their cell wall type: gram_negative, gram_positive, gram_variable¶

Chapter 2

Methods

The dataset consisted of 108 metabolic pathways of living organisms' kingdom details- 77 Bacteria, 2 Chromista, 7 Animal, 2 Fungi, and 6 Protozoa. All organisms classified by the kingdom which they belong to, from Cavalier-Smith six-kingdom models as below:

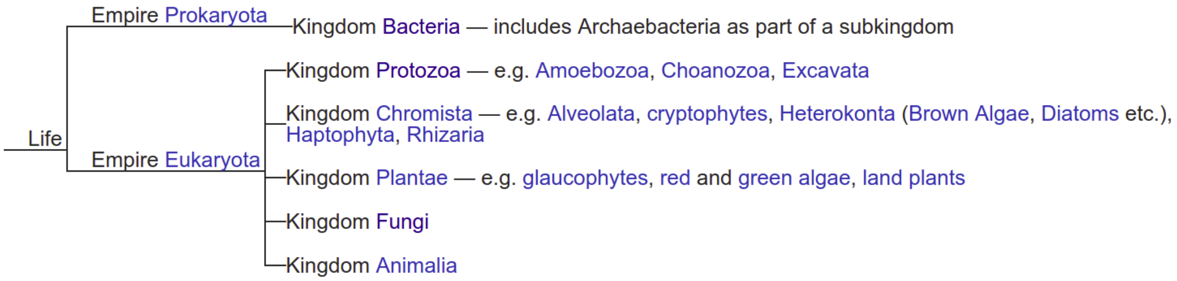

Life
Empire Prokaryota
Kingdom Bacteria — includes Archaebacteria as part of a subkingdom
Empire Eukaryota
Kingdom Protozoa — e.g. Amoebozoa, Choanozoa, Excavata
Kingdom Chromista — e.g. Alveolata, cryptophytes, Heterokonta (Brown Algae, Diatoms etc.), Haptophyta, Rhizaria
Kingdom Plantae — e.g. glaucophytes, red and green algae, land plants
Kingdom Fungi
Kingdom Animalia

Figure 3: Figure shows the kingdoms of all the 108 organisms

The models obtained from the BiGG platform needs to be processed and prepared for the clustering technique. In order to prepare the data, the models were convereted to a Directed Graph (DAG) using the cameo package of python. The DAG was then converted to Embedded graphs where the weight, standard deviation and skewness could be studied and added as a feature set for the hierarchical clustering.

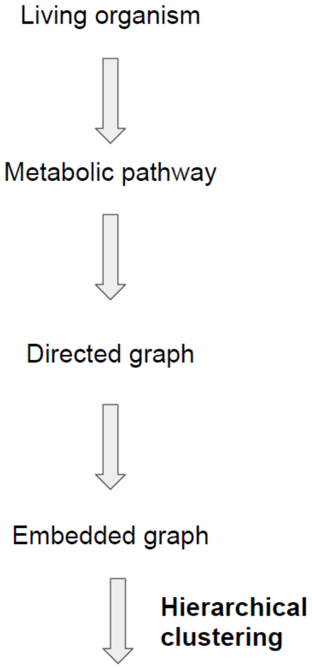


Figure 4: Flowchart of the methodology followed during the project

.

2.1 Directed Graphs

The directed graphs were created using the cameo package’s network analysis and mapping feature. Since 5 categories of the organism kingdoms were identified, a mapping was created and visualized.

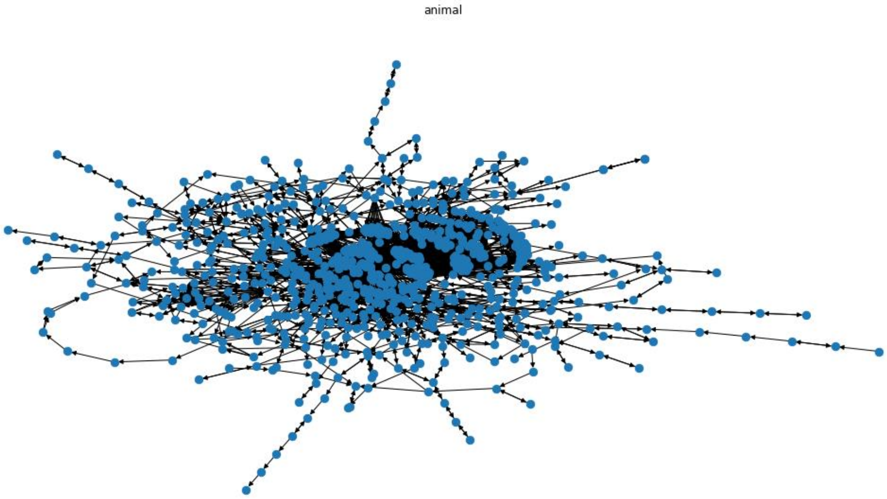


Figure 5: Network Mapping of Animal - Homo sapiens - iAT_PLT_636; Number of nodes: 737; Number of edges: 2423; Average in degree: 3.287; Average out degree: 3.2877; Average shortest path: 1.1; Average clustering coef: 0.145

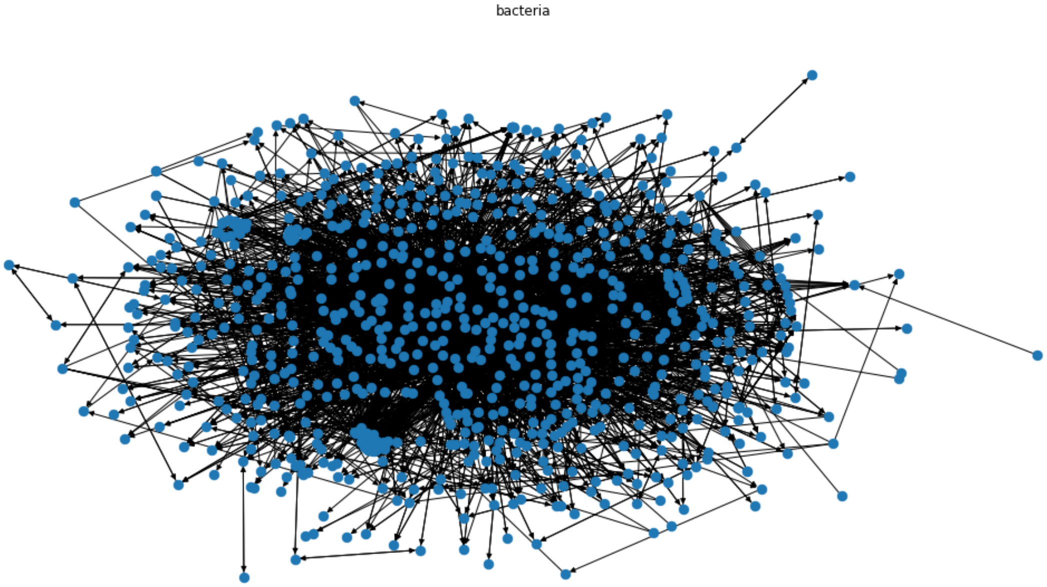


Figure 6: Network Mapping of Bacteria - Acinetobacter baumannii AYE - iCN718; Number of nodes: 851; Number of edges: 4382; Average in degree: 5.1492; Average out degree: 5.1492; Average shortest path: 0.938; Average clustering coef: 0.202

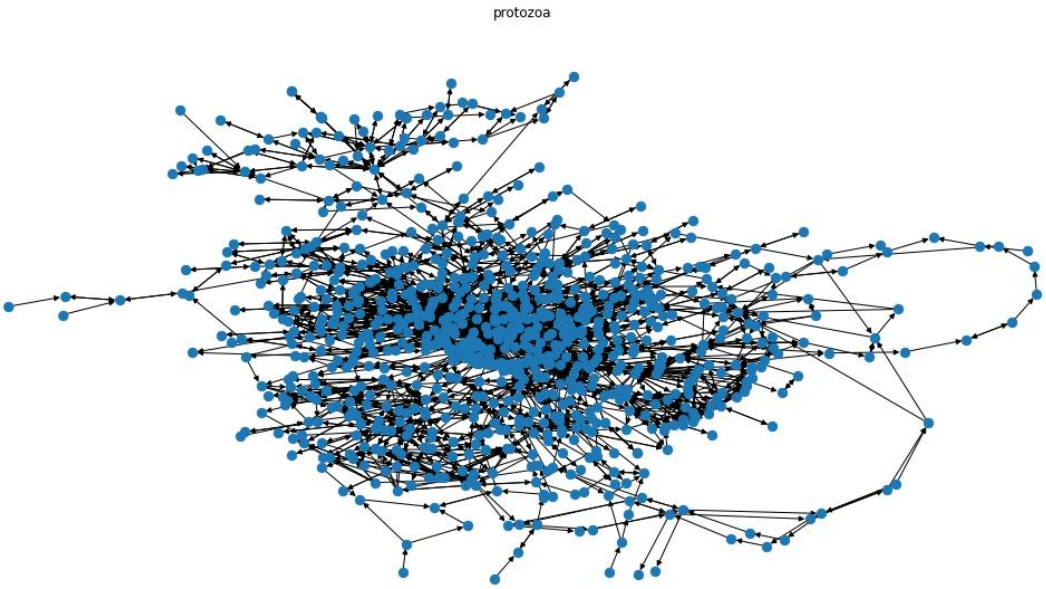


Figure 7: Network Mapping of Protozoa - Plasmodium vivax Sal-1 - iAM_Pv461; Number of nodes: 896; Number of edges: 2251; Average in degree: 2.5123; Average out degree: 2.5123; Average shortest path: 0.994;Average clustering coef: 0.139

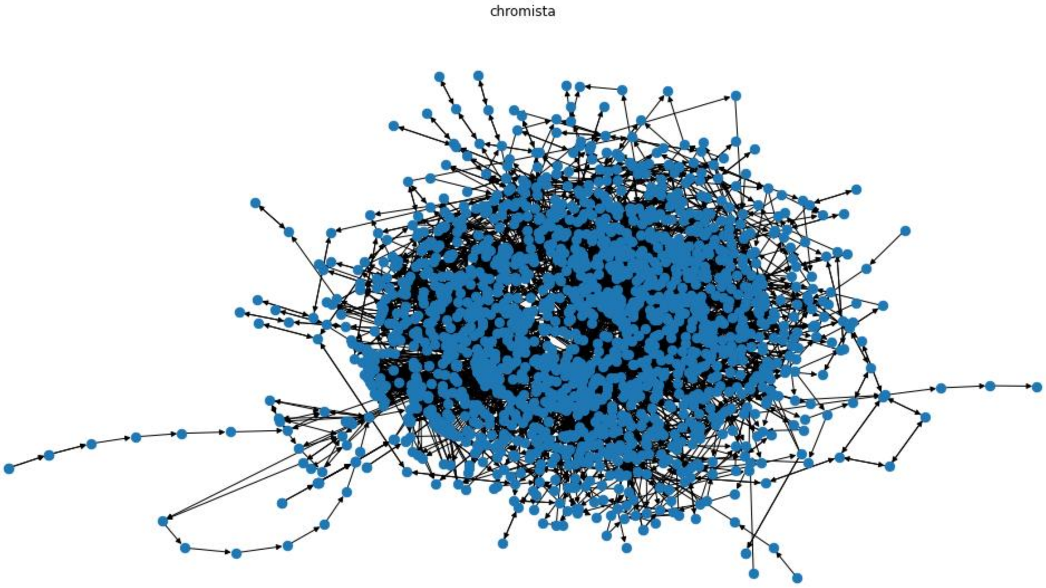


Figure 8: Network Mapping of Chromista - Chlamydomonas reinhardtii - iRC1080; Number of nodes: 1701;Number of edges: 5868;Average in degree: 3.4497;Average out degree: 3.4497;Average shortest path: 1.131;Average clustering coef: 0.168

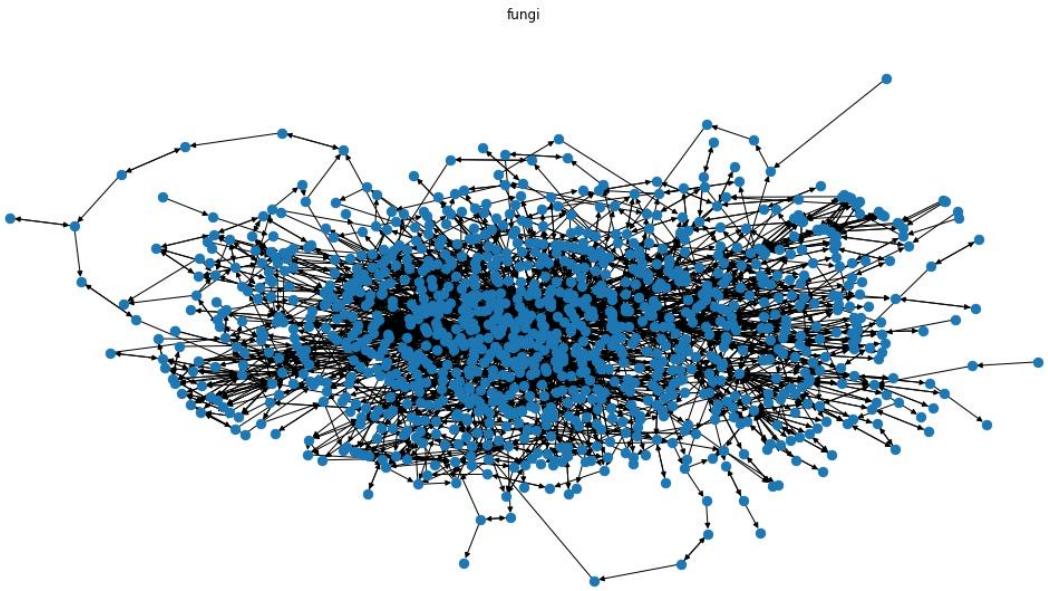
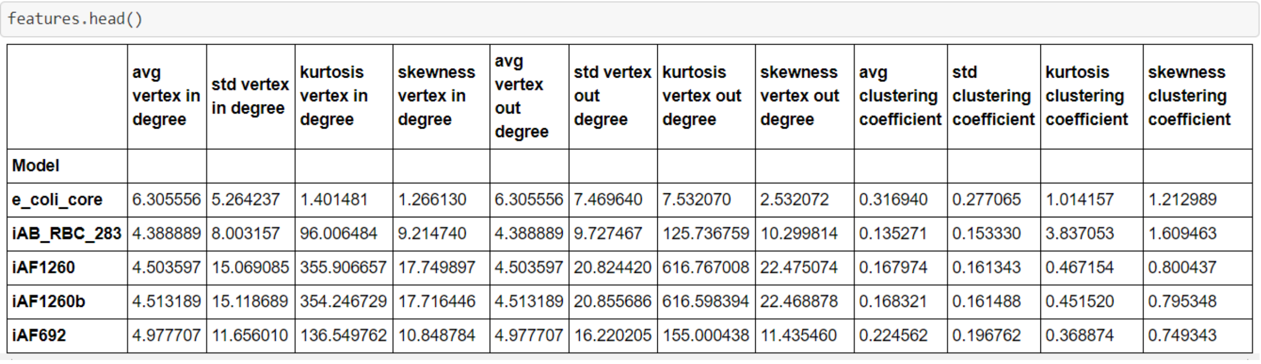


Figure 9: Network Mapping of Fungi - Saccharomyces cerevisiae S288C - iMM904; Number of nodes: 1170; Number of edges: 3207; Average in degree: 2.7410; Average out degree: 2.7410; Average shortest path: 1.121; Average clustering coef: 0.156

2.2 The Graph Embeddings

The graph embeddings are necessary to convert the directed graphs into a feature vector to help the machine learning model to cluster the similar groups based on these features. The calculation of features like – standard deviation, average, kurtosis, and skewness of the directed graphs each node’s indegree vertex, outdegree vertex and clustering coefficient was created. Kurtosis is a factual measure that characterizes how intensely the tails of a distribution contrast from the tails of an ordinary appropriation. As such, kurtosis distinguishes whether the tails of a given distribution contain outliers. The standard deviation and mean were taken to calculate and observe the distribution of the entity compared to the whole dataset.

The features were stores in a embedded_grahs.csv file in order to call back to the clustering algorithm.



```
features.head()
```

	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
Model												
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

Figure 10: Feature Vectors of each of the DAG

The features were smoothened by scaling it to a comparative value under a distribution. This is also done to avoid bias into the machine learning algorithm and reduce the noise against the extreme outliers in the data points. The formula used was:

features_scaled = (features - features.mean()) / features.std()

2.3 The Hierarchical Clustering

A Hierarchical Clustering method, implemented using the sklearn package Agglomerative Clustering, was used to create a clusters. A dendrogram was first constructed to understand the need for the number of clusters and the hierarchical ordering was analyzed for the hierarchical clustering.

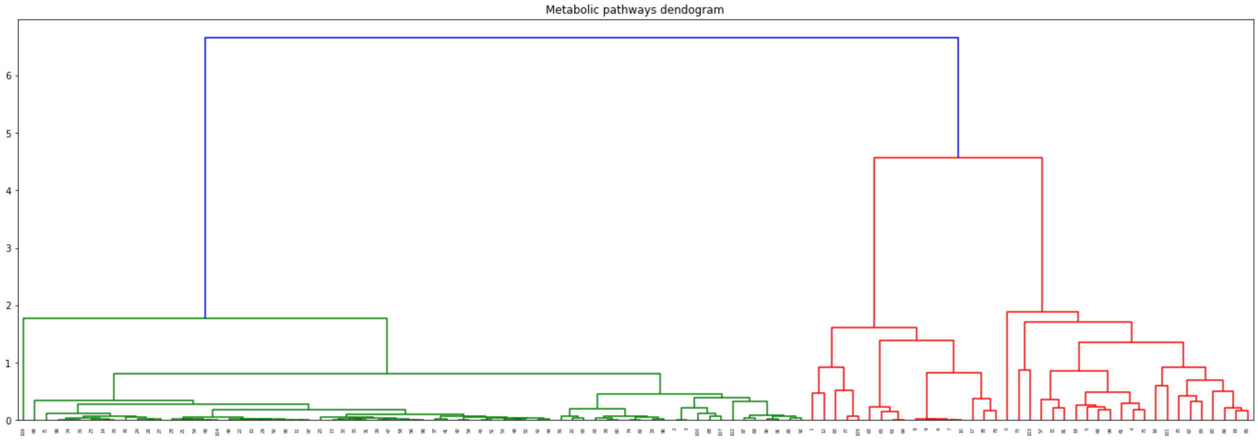


Figure 11: Dendrogram based on kingdoms to find number of clusters to be interested in

The start of the experiment was the simple definition of 2 clusters where the dataset was clustered into 2 groups of empires based on their metabolic pathways. We observed that for Eukaryotes: 19 out of all 20 eukaryote organisms where correctly put into cluster 1 and for Prokaryotes: 68 out of all 88 prokaryote organisms where correctly put into cluster 0

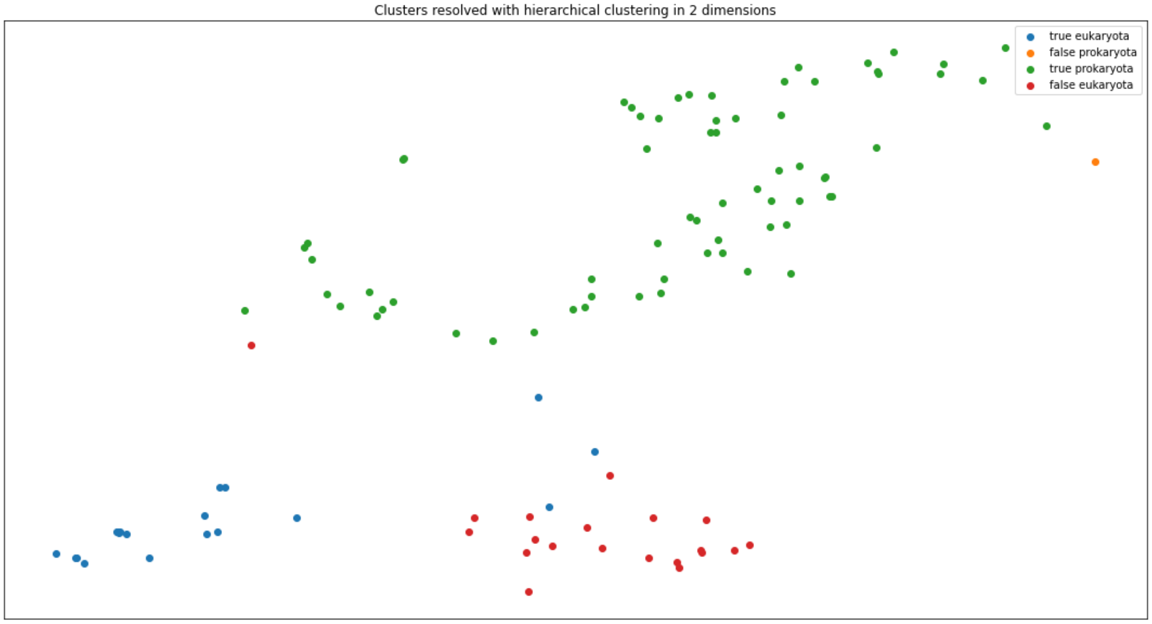


Figure 12: Reduced the feature dimension to 2 dimension using tSNE for 2 clusters

The second experiment was done with 5 clusters of the kingdoms that the organisms belong to. We can see that hierarchical clustering was able to group organisms into 5 kingdoms based on their metabolic pathways a bit worse than into empires but still well. For Animalia: 4 out of all 7 animalia organisms where correctly put into cluster 6. For Bacteria: 85 out of all 88 bacteria organisms where correctly put into cluster 1 or 2. For Chromista: 1 out of all 2 chromista organisms where correctly put into cluster 5. For Fungi: 2 out of all 2 fungi organisms where incorrectly put into cluster 0 and should be in cluster 3. For Protozoa: 9 out of all 9 protozoa organisms where correctly put into cluster 0

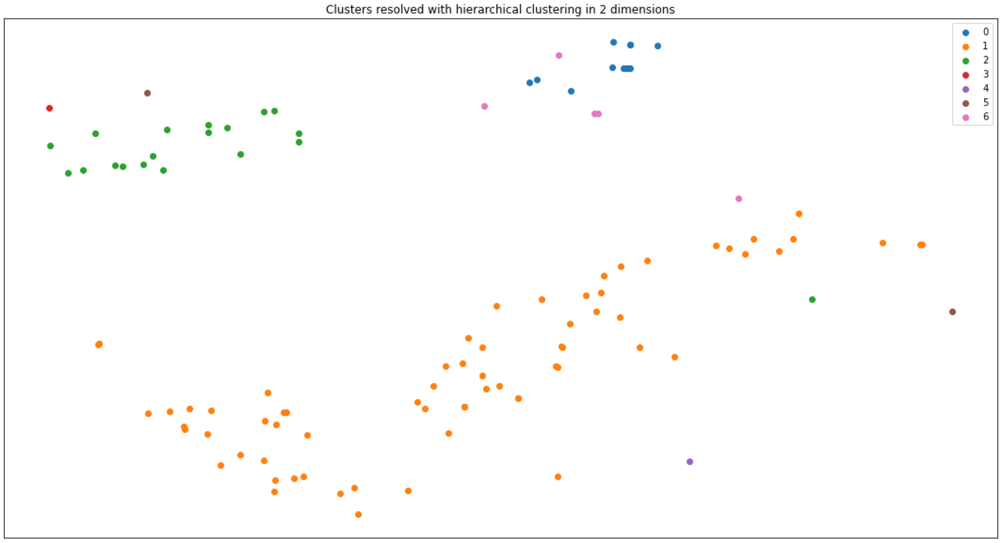


Figure 13: Reduced the feature dimension to 2 dimension using tSNE for 5 clusters

The final experiment was constructed around the types of the cell walls- gram positive, gram negative and gram variable (as miscellaneous). A dendrogram was constructed to identify the number of clusters requirement:

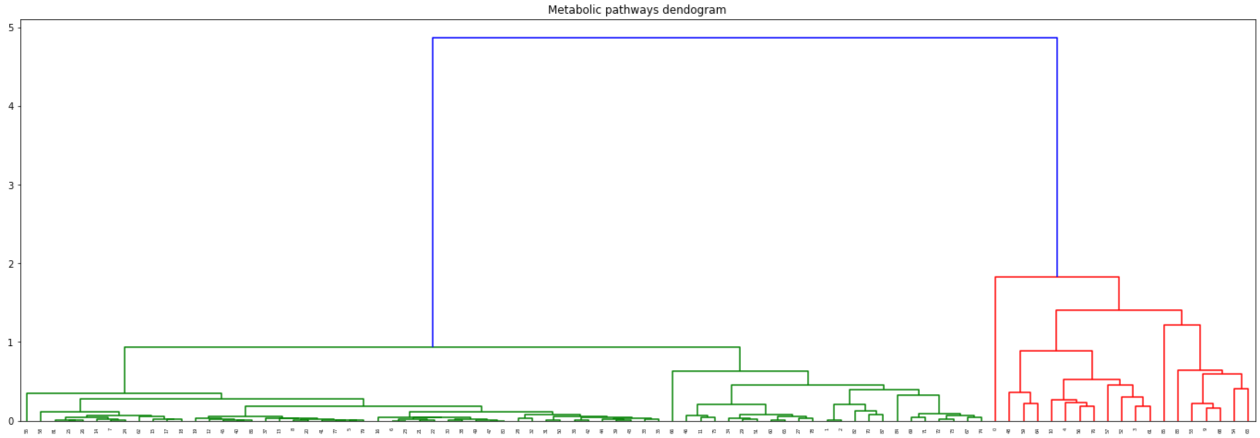


Figure 14: Dendrogram based on cell wall type to find number of clusters to be interested in

After the hierarchical clustering into 3 clusters, the observation was:

gram-negative: 69 out of all 79 gram-negative bacteria where correctly put into cluster 1

gram-positive: 8 out of all 8 gram-positive bacteria where correctly put into cluster 0

gram variable: 1 out of all 1-gram variable bacteria where incorrectly put into cluster 0 and should be put into cluster 2

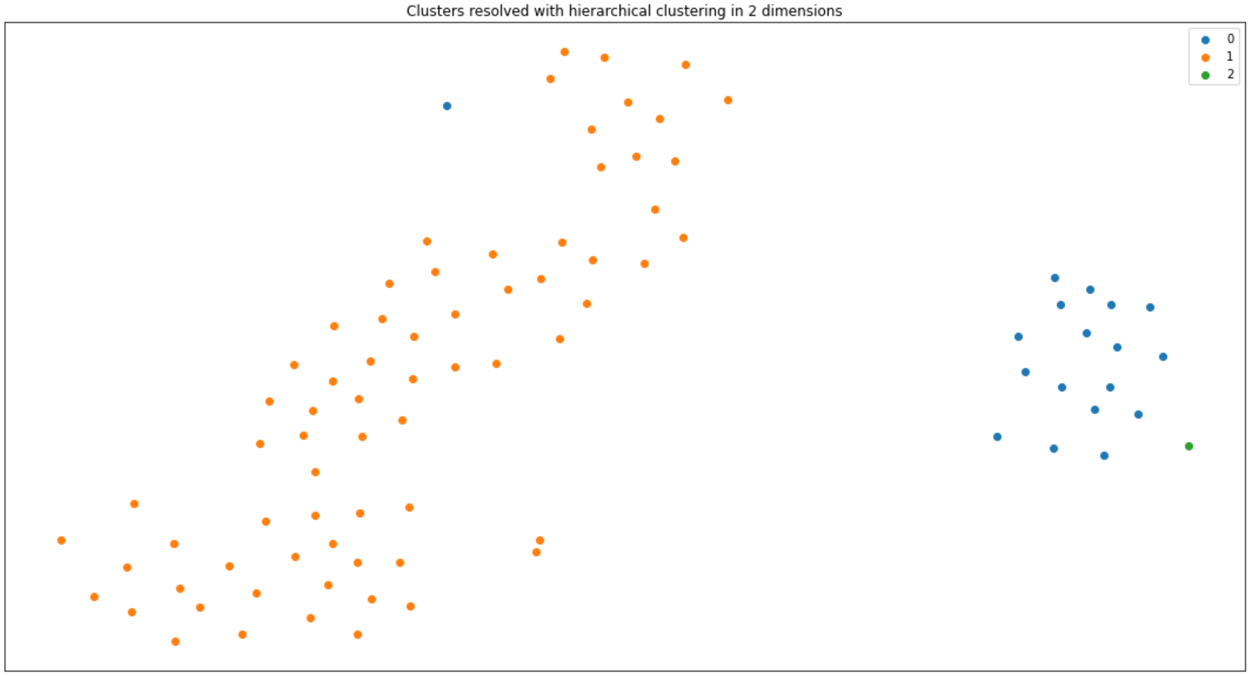


Figure 15: Reduced the feature dimension to 2 dimension using tSNE for 3 clusters

Chapter 3

Results

3.1 Understanding the Clusters

The construction of the bacterial phylogenetic tree by Jonathan Eisen [15] was used as a reference guide to help understand the clustering hierarchy of the organisms in the dataset.

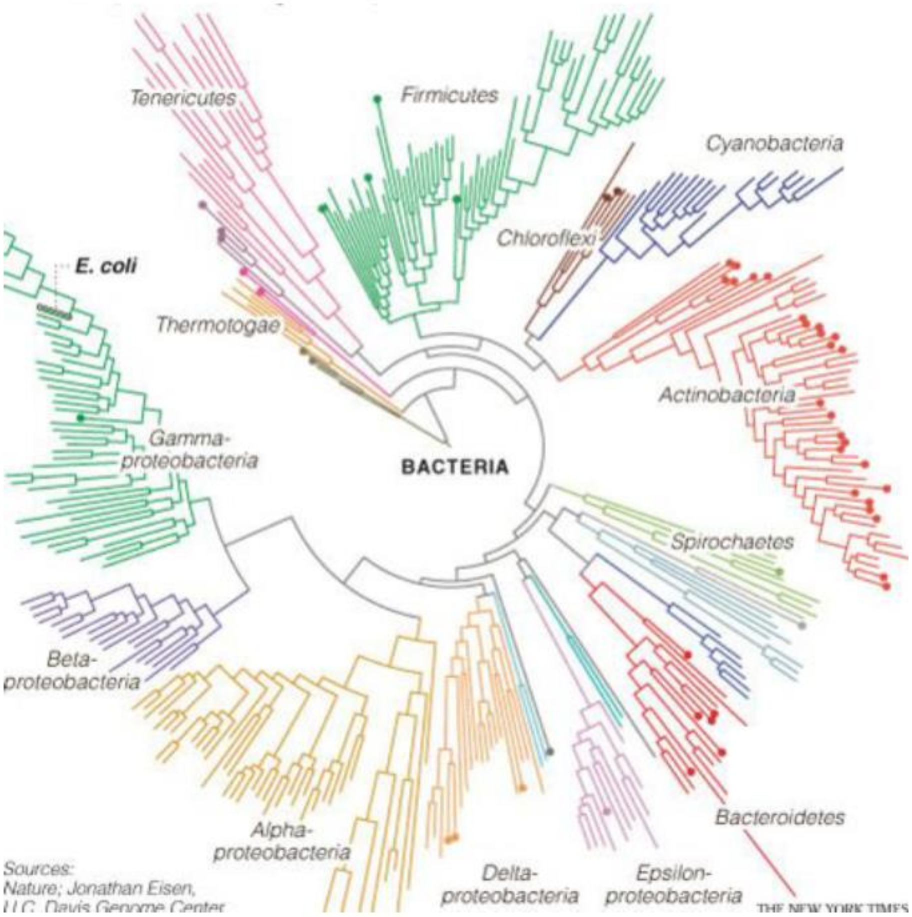


Figure 16: The Bacterial Phylogenetic Tree [15]

3.2 Genus- Escherichia

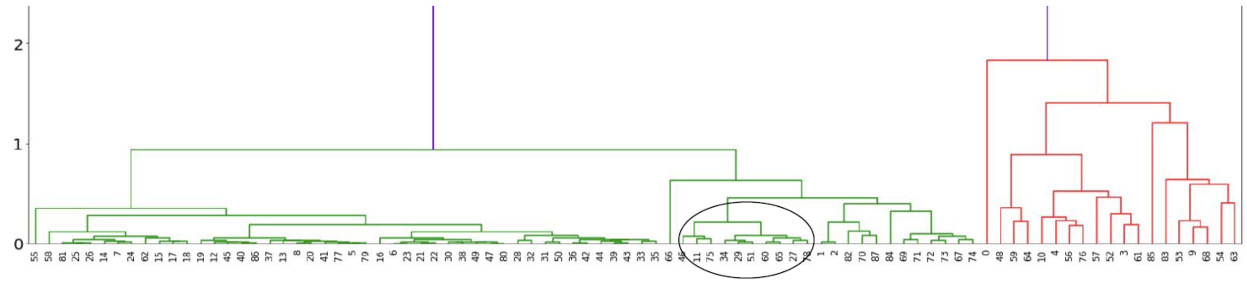
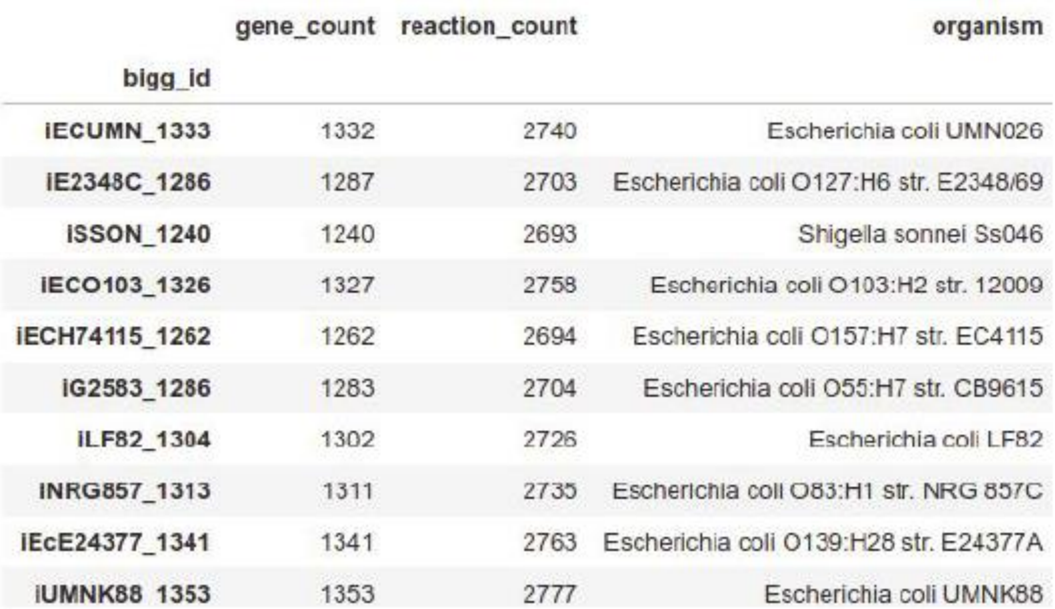


Figure 17: 9 out of 10 bacteria from that cluster are indeed from genus Escherichia

On comparing with the original dataset the results were confirmed:



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

Figure 18: Dataset of Escherichia from BiGG models

Also all of the bacteria in this cluster marked in Figure 19 are in phylum - Proteobacteria all the way down in hierarchy to genus - Escherichia

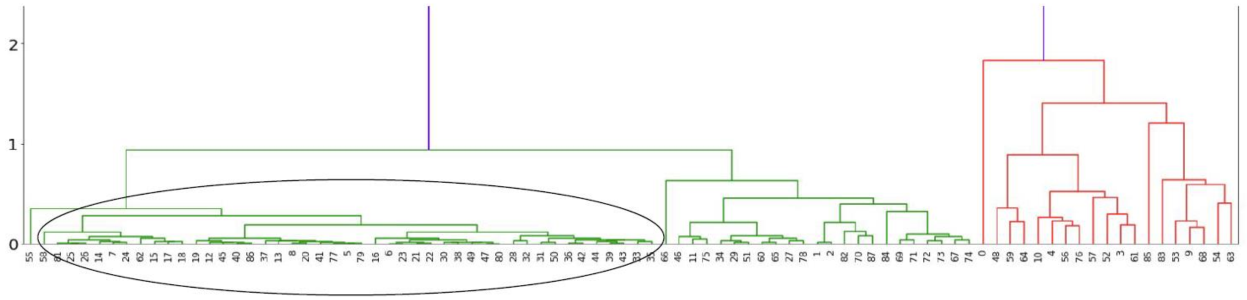


Figure 19: Phylum Proteobacteria cluster

There are some Bacteria in the cluster can’t be appropriately separated on any level up to phyla where bacteria belong to phylum: Proteobacteria, Actinobacteria, Firmicutes, Cyanobacteria, Thermotogae

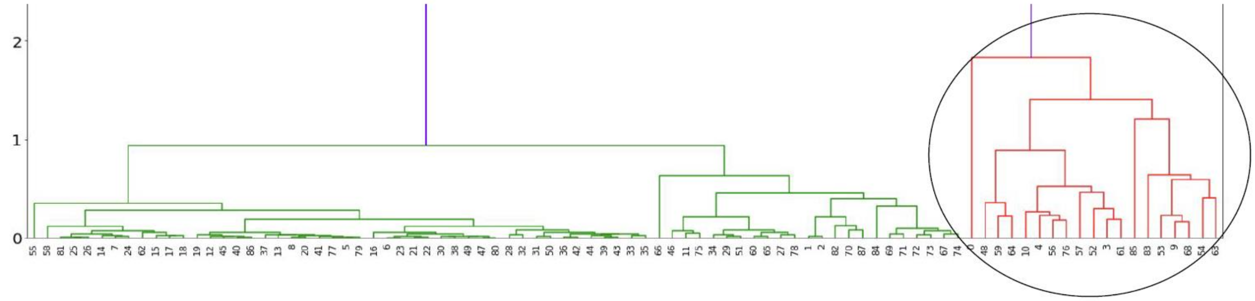


Figure 20: Unclear Phylum

Chapter 4

Discussion

4.1 Limitations

This method requires a creation of embed graphs called feature vectors in order to form groups based on similar attributes. There can be more statistical methods applied for generating this vector set. The standard deviation, the mean and the kurtosis are the only standard benchmark techniques used here. More the feature set, the better the clustering algorithm would predict the clusters.

A further limitation for scaling the feature set and having a well-defined feature engineering technique is needed. The Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model. In this approach the usage was limited to only Min-Max Scalar which just averages based on the outlier values. Hence a better approach for feature scaling could be used like a Absolute Maximum Scalar and Unit Vector Scaler that helps the outliers to be smoothened for avoiding high variances in the dataset.

4.2 Future Work

The fact that the method relies on a training set means that roughly some of the phylum of the outcome variable are still not concluded in the clusters due to the lack of adequate training data sets and proper feature engineering. One remaining step of this project, therefore, is to expand the development to including different clustering techniques- Density Based Clustering (DBSCAN) or an Ensemble Clustering technique involving multiple clustering methods. These categories are likely rare within the dataset and thus they could not be located by searching the dataset for keywords related to the phylum. It is possible for some of these categories there are not enough datasets deposited in the dataset at all and that the amount of required data to build a classifier is not publicly available at this time.

In addition, to expanding the training set to include the currently unclear categories, more samples from the minority categories (i.e. those that do not yet contain more data points in the training set) should be included in order to achieve balanced input data, which may increase the accuracy of these categories.

Finally, a more balanced representation of the bacteria from each source within the set of bacterial genomes from the dataset that are used to create the input features may also lead to improved accuracy. Bacteria that have been more frequently studied like the phylum Escherichia, have a higher representation in this database and accordingly, more of these categories were annotated for inclusion.

4.3 Conclusion

In the Figure 21 clusters with its members genus – Escherichia (blue) were identified. Also, cluster with its members genus - Shigella (black) was identified. For the bacteria on the red branch the clusters with its members could not be identified that belong to same group on any level in Linnaean Taxonomy (red). Additionally, the clustering with avg. shortest path feature was performed but the results did not improve compared to the baseline

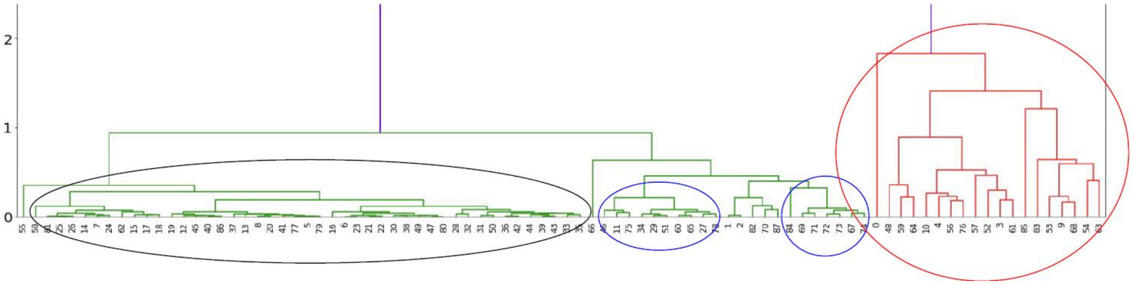


Figure 21: Clusters Identification

Appendix A

Table of Figures:

References

1. https://bhavanajagat.com/2014/06/19/spirituality-science-the-human-species/
2. Dinsdale, E. et al. Functional metagenomic profiling of nine biomes. Nature 452, 629–632 (2008)
3. Alberich R, Castro JA, Llabrés M, Palmer-Rodríguez P. Metabolomics analysis: Finding out metabolic building blocks [published correction appears in PLoS One. 2017 Oct 12;12 (10):e0186626]. PLoS One. 2017;12(5):e0177031. Published 2017 May 11. doi:10.1371/journal.pone.0177031
4. Dandekar, T., Schuster, S., Snel, B., Huynen, M., and Bork, P., Pathway alignment: application of the comparative analysis of glycolytic enzymes. Biochem. J 343(1999)115-124.
5. Durbin, R., Eddy, S., Krough, A., and Mitchison, G., Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998.
6. Felsenstein, J. 1993. PHYLIP (Phylogenetic Inference Package), http://evolution.genetics.washington.edu/phylip/software.html
7. Forst, Christian V., and Schulten, Klaus, Evolution of Metabolisms: A New Method for the
8. Comparison of Metabolic Pathways Using Genomics Information. Journal of Computational Biology 6(1999)343.
9. Lin, Jimmy and Gerstein, Mark, Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for comparing Genomes on Different Levels Genome Research 10(2000)808-818.
10. Overbeek, R., Niels, L., et al, WIT:integrated systems for high throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Research 28(2000)123-125.
11. Page, R.D. TreeView: An application to display phylogenetic trees on personal computers. Comput.Appl. Biosci 12(1996)307-331.
12. Tatusov, R.L., Koonin, E.V., and Lipman, D.J.. A genomic perspective on protein families. Science 278(1997)631-7; Tatusov, R.L., et al The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29(2001)22-8
13. Wang, Jason T.L., and Zhang, Kaizhong, Finding similar consensus between trees: An algorithm and a distance hierarchy. Pattern Recognition 34(2001)35-45.

14. Zhang, S., Liao., L., Tomb, J-F., and Wang, Jason T.L., Clustering of metabolic pathway data using different tree alignment distances. Submitted to BioKDD 2002
15. Wu, D., Hugenholtz, P., Mavromatis, K. et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462, 1056–1060 (2009). https://doi.org/10.1038/nature08656
16. Cameo Package Python- https://pypi.org/project/cameo/