Bacterial Metabolic Networks Analysis

Group 10

Dylan Smith

Description

For my data mining project, I want to build and compare bacterial metabolic network models. Some of the interesting questions that can be answered with this is investigating how different organisms accomplish similar metabolic tasks (lipid metabolism, carbohydrate metabolism, etc), whether tasks are conserved across organisms, how much of their genome they devote to different metabolic tasks, etc.

Prior Work

Functional Comparison of metabolic networks across species

- https://www.nature.com/articles/s41467-023-37429-5

Predicting metabolic modules in incomplete bacterial genomes with MEtaPathPredict

- https://elifesciences.org/articles/85749

Datasets

The dataset I'll be using for this project is the KEGG (Kyoto Encyclopedia of Genes and GEnomes) database. KEGG is a common dataset in bioinformatics and contains molecular-level information for a large number of organisms obtained through genome sequencing and other high-throughput methods. It includes data for different networks, genes, genomes, reactions, enzymes and more. For this project, I'll be focused on the Networks, reactions, and enzymes datasets for different bacteria and may bring in the genes dataset as well.

These datasets are accessed through a REST API, so I can get the relevant information without having to download the entirety of the information they have on each organism. I have also found a package KEGGutils which integrates the KEGG API with NetworkX which will be useful in a project analyzing networks.

- KEGG: Kyoto Encyclopedia of Genes and Genomes
- KEGGutils/tutorials/Tutorial 1 - EnzymeGraphs.ipynb at master · filippocastelli/KEGGutils (github.com)

Proposed Work

The data from KEGG is accessed through the REST API rather than from direct downloads, so a big part of this project will be accessing and formatting all of the relevant data through the API. The data is interconnected in their database, so I will have to write a script to access the data for each organism, get the relevant enzymes, genes and reactions for the different metabolic pathways.

This database is well established and maintained so the value formats are standardized and for the most part clean. Since this project is a comparison across species, the list of species accessed can be chosen to include more studied and common organisms so the likelihood of missing data is decreased. For those that aren’t full, there are perhaps tools that can be used to predict the function of unknown genes or enzymes.

Most of this will entail building functions that access the metabolic network data for different organisms, accessing and acquiring the corresponding genes, enzymes, metabolites and reactions, and then putting that in a format where the metabolic models can be built.

KEGG breaks the pathways down by function, but it might be possible to incorporate them into a whole network to see the interconnections between them.

Evaluation

This project is on the exploratory side so an evaluation will entail more of whether the models built are coherent. In order to compare the models for each of the organisms, I can compare the components of the functional sections between organisms to see how different organisms achieve similar function. I can compare different network metrics such as connectivity to see how the network aspects differ between organisms. Relative size devoted to different functions can also be compared across species to see how different bacteria distribute their overall metabolism to different functions.