Bacterial Metabolic Networks Analysis

Dylan Smith
CSCI 4502
University of Colorado, Boulder
Boulder Colorado, USA
dysm3436@colorado.edu

PROBLEM STATEMENT

Despite the wealth of genomic data available on a wide array of organisms, systems biology and synthetic biology are still hindered by our limited understanding of how biological systems arise and function. Metabolic networks are a prime example of our limited understanding, despite forming the basis for much of cellular activity we are still lacking a deep understanding of the universal principles that govern how these networks operate as well as how they change and adapt to different environments.

While some metabolic networks are conserved across species, a defining trait of organismal complexity is the increase in overall size of metabolic network complexity. Thus, for both the conserved metabolic modules as well as the limited overall metabolic network complexity, bacteria provide a simpler model to study universal principles of metabolic networks. Including how different organisms accomplish similar metabolic tasks, and the extent to which these tasks are conserved across species.

This project aims to build and compare bacterial metabolic network models to investigate these questions, focusing on key metabolic pathways such as lipid metabolism and carbohydrate metabolism. By analyzing these networks, we can explore how bacteria allocate their genomic resources to various metabolic functions and identify patterns of conservation and divergence among different species. We can also explore the properties of these networks to explore how the structure and connectivity of similar networks varies across species.

In summary, this project aims to build and compare bacterial metabolic network models to uncover underlying principles that govern and control metabolism. Developing this understanding could have significant implications in understanding biology and evolution, and could lead to application in synthetic biology and genetic engineering.

LITERATURE SURVEY

In the paper Functional comparison of metabolic networks across species, Ramon and Stelling address the challenge of understanding how evolutionary history and environmental adaptation shape metabolic phenotypes in microbes. The authors propose a method of linking genotype and environment to phenotype using sensitivity correlations to compare metabolic network responses to perturbations, allowing for a detailed comparative analysis of bacterial metabolic networks. By identifying conserved and variable metabolic functions across 245 bacterial species, the paper provides insights into which metabolic tasks are conserved and which are variable, and provides a structure for this project.

In the paper Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict, the authors describe MetaPathPredict which uses deep learning to predict the presence of metabolic pathways (KEGG modules) in bacterial genomes. By accurately reconstructing metabolic pathways, MetaPathPredict enables understanding of functional properties of bacterial metabolic networks, and allows for a more robust comparison of metabolic networks by determining functional modules that can be further compared.

PROPOSED WORK

The dataset for this project will be sourced from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database which is a well-established and commonly used resource in bioinformatics. The dataset contains the relevant data on Networks, Reactions, Enzymes and Genes, and can be accessed through the KEGG REST API to allow for selective retrieval of the relevant information without having to download entire datasets. The KEGGutils package integrates this KEGG API with NetworkX to facilitate the network analysis needed for this project.

The KEGG database structures data across species in a consistent format, ensuring clean data that is compatible for comparison across species. The database contains data for thousands of species, with varying levels of research interest leading to varying levels of data integrity. Focus on more studied and common bacteria reduces the likelihood of missing data and allows for bacterial species with well-defined and clean data relevant to metabolic networks to be chosen for analysis. These steps and the established format used, ensures that little data preprocessing is needed with this data.

In order to use this data to build metabolic networks for the different bacterial species, scripts will need to be written to access and retrieve the relevant data for each organism. This script can then be used to retrieve the data for each of the list of organisms, and can be coupled to a script that will build the metabolic networks to be used for further analysis. The data from the KEGG database has an interconnected nature which allows for associated data to be retrieved based on previously known data. The functions for this retrieval can thus be written such that the data for each organism is systematically obtained to ensure that the complete metabolic network can be built.

For well studied organisms, the metabolic pathways are already broken down into modules based on metabolic function, so choosing appropriately well-studied organisms allows us to build and tag metabolic networks based on function. In the case this is limited or for any further analysis beyond the scope of this project, predictive tools such as those mentioned in the previous work could possible be used to build metabolic networks based on functionality for less well-studied organisms.

Once the metabolic networks are built for each of the organisms, a comparison and analysis of the networks across different species can be carried out. Network modules and their relative sizes can be compared to analyze how networks are allocated according to function, which could elucidate where function is more likely to be conserved or varied. Different network properties such as connectivity, centrality, betweenness, closeness, density and other properties can also be defined and compared to analyze how the network structure of different functional modules differs across species and whether those changes correlate to the pieces that make up the different networks.

DATASET

As mentioned in the previous section, the dataset to be used for this project is the KEGG (Kyoto Encyclopedia of Genes and Genomes) dataset.

https://www.kegg.jp/kegg/

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.

EVALUATION METHODS

This project is primarily exploratory, focusing on the coherence and validity of the constructed metabolic models Thus the evaluation will involve assessing the internal consistency and biological consistency of each bacterial metabolic network model. To that extent, the models created can be compared to those described in the KEGG database to ensure all of the data is properly incorporated.

For a comparative analysis of the metabolic networks across the species, we will undertake the following:

Functional Component Comparison: We will compare the functional modules of the metabolic networks across organisms to analyze how different species achieve similar metabolic functions and how they allocate resources across function.

Network Metrics Analysis: In order to understand the structural differences between the metabolic networks of the different species, we will analyze different network metrics such as connectivity, degree centrality, clustering coefficients and others. This comparison will help elucidate how the properties of the networks vary across species and how these metrics change with other aspects of the metabolic networks.

TOOLS

For this project, the data will be accessed using the KEGG REST API and the KEGGutils package. Network construction and analysis will be carried out using the NetworkX python package.

MILESTONES

The largest and most complex part of this project is writing the script to access the relevant information from KEGG needed to reconstruct the metabolic networks. The following comparative analysis is relatively easy, so to stay on track for this project, I want to have the script to access and build the models finished in the next two weeks. The script should be the same for the different species, so I need to have it finished and working for one by that time, and then I can just run it on the other organisms chosen. Depending on the computational drain of the process, I want to have models built for at least 10 different bacteria.

The following week I want to spend writing the script to get the metrics for the comparison analysis between the organisms which will include getting the relevant size of the functional modules as well as getting the network properties of each network. Following a similar style to the first section, a script can be written for the first organism that can then be run on all of the others, with the results being stored away for side by side comparison and analysis.

This should leave enough time for the analysis and writeup of the project in the remaining time.

MILESTONES COMPLETED

There are 5 main components to the Milestones I have for this project:

1. Write a script using KEGGutils in python to access and acquire all relevant information from KEGG database including
    a. Networks
    b. Reactions
    c. Enzymes
    d. metabolites
2. Write script using NetworkX in python to reconstruct and visualize the metabolic network
3. Apply the two previous scripts to at least 10 organisms (depending on computation load)
4. Perform comparative analysis between the metabolic networks of the organisms from milestone 3

The bulk of this project consists of step one and two above as once the script to extract and build a metabolic network for a single organism is completed, it can be run on other bacterial organisms with the only limit being computational load. Milestone 4 should also be relatively simple as NetworkX has many prebuilt functions for network analysis.

So far the only milestone that I have essentially completed is milestone 1. My script currently takes in the organism abbreviation and creates a dictionary that establishes bacteria-enzyme relations and enzyme-metabolite pair relations. This is enough to create a metabolic network for the organism of interest, but I don't currently have a method to differentiate which subnetwork each enzyme is

associated with which will be important for comparative analysis relating to the distribution of resources towards the different subnetworks. It is easier with the way KEGG is setup to go from organism -> enzyme -> reaction -> metabolites than it is to go from organisms -> met pathway -> enzymes -> reaction -> metabolites. Missing is that met pathway step in the middle, although since certain enzymes or metabolites may participate in multiple metabolic subnetworks, it may be easier to tag each enzyme and reaction to the subnetwork rather than the other way around.

MILESTONES TODO

From the milestones listed above, there are still milestones 2-4 still to do as well as finishing milestone 1 with tagging each enzyme and reaction with the sub metabolic network (s) it is associated with. Steps 2 should be fairly straightforward as the way I have setup step 1 facilitates the use of NetworkX for the construction of a Bipartite network (bipartite because there will be two types of nodes. Enzymes and metabolites). The output from milestone 1 is a dictionary with the reaction as the key and the value containing the enzyme, metabolites (inputs and outputs), and associated subnetworks.

With that, the construction of the network using NetworkX can be performed by iterating through the reaction network to create nodes for the enzymes and metabolites and edges if they are in the same reaction. They will be directed edges where it will go from metabolite node to enzyme node if the metabolite is an input, or from enzyme node to metabolite node if the metabolite is an output. The associated subnetwork can be used to tag the enzymes and metabolites that are involved in that specific sub network, which will allow for conditional distillation of the entire network based on each subnetwork.

Alternatively, I may build the metabolic network by subnetwork first so that the subnetworks are separate from each other and thus somewhat pre-clustered. This could help with visualization purposes as I'm not currently sure if NetworkX has the capability to single out subnetworks based on tagged information.

An aspect that I will have to consider is that while the reactions, enzymes and metabolites are likely known for the organisms that will be chosen, the specific function or subnetwork they are associated with may not be known to the same degree. Thus I may have to include an unknown tag for these entries or perform some other method to predict or fill out those entries.

For milestone 3, I will just have to choose the organisms to include in the project. I'll focus primarily on well studied organisms as they are the most likely to have fully studied and complete metabolic networks. This information is relatively easy to find as there are a relatively few number of well studied organisms and KEGG has information pertaining to how filled out the information for each organism is.

Milestone 4 will comprise most of the analysis that hopefully makes this project interesting. I could now find any previous studies that compared metabolic networks based on their network structure. While the number of organisms that will be studied here is not comprehensive by any means, I hope that this comparison can provide insights as to how metabolic networks are evolved, maintained and distributed across species.

Some of the metrics that I will measure and compare are:

Individual node metrics: These metrics are determined for each node in the network and measure the influence of that node within the network.

- Degree Centrality: measures the number of connections (edges) a node has. Nodes with high degree centrality are considered influential within the network.
- Betweenness Centrality: measures the extent to which a node lies on paths between other nodes. Nodes with high betweenness centrality are involved in flow throughout the network.

- Closeness Centrality: measures the average length of the shortest path from a node to all other nodes in the network. Nodes with high closeness centrality can quickly interact with other nodes and are efficient at spreading flow throughout the network.
- Eigenvector Centrality: measure the influence of a node based on the influence (connections) of it's neighbors.

Network metrics: these metrics measure aspects of the network as a whole and can be used to compare different networks.

- Clustering Coefficient: measure the degree in which nodes in a network tend to cluster together. High clustering networks indicate redundancy in a network.
- Average Path Length: measure the average number of steps along the shortest paths for all possible pairs of network nodes. Short average path lengths suggests that any node can be reached from any other node which is efficient for flow through the network.
- Density: measures the proportion of potential connections in a network that are actual connections and is a measure of interconnectedness.

The above are all common metrics used to measure different aspects of a network. The node metrics can be graphed and compared to see how different nodes and reactions compare across the different organisms. Many reactions are shared between organisms, so comparing the importance of those shared reactions can possibly provide insights into those reactions and their role in each of the organisms.

The network metrics above can be used to compare how network properties vary across the different organisms. This can tell us if these network properties are conserved or different among organisms. If they are different or similar within subnetworks.

Much of these metrics are measures on how flows through a network occur. These flow considerations are things that can only be seen when looking at the network as a whole and how its connected, and have possibly novel implications towards how biology works.

RESULTS SO FAR

I am currently just building the logic to extract the information and to build the network so there is little to report yet as far as results.