Bacterial Metabolic Networks Analysis

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

ABSTRACT

Understanding the principles that govern the complexity inherent in metabolic networks is a crucial step in advancing systems and synthetic biology. Although an extensive amount of genomic data is available for a wide range of organisms, the complexity of biological systems remains a significant challenge. Complexity in general is a nascent field and much research is still needed to understand even the most simple complex systems, much less biological systems. Metabolic networks are one such complex biological system where our limited understanding of biological complexity is on display. This study aims to contribute to this gap by building bacterial metabolic network models and focusing comparative analysis on network properties rather than individual components of the reaction. By examining and comparing networks across different bacterial species, we seek to uncover patterns of conservation and divergence that shed light on how organisms allocate genomic resources to metabolic functions.

Our analysis found that each of the ten bacteria had similar trends across the sub metabolic networks we investigated, indicating that allotment of resources to subnetworks (measured by the number of nodes) is somewhat conserved by function. Network properties such as average path length also varied more so by function than by species indicating some amount of conservation of network structure across species as well.

INTRODUCTION

Despite the rapidly expanding repository of genomic and other biological data, our comprehension of biological systems, including metabolic networks, remains incomplete. Metabolic networks are a key foundation to cellular activity, yet the principles that govern their function and evolution are not well understood. While portions of these networks are well conserved across species, much of them contain significant variation, particularly as organismal size and complexity increases. Bacteria, with relatively simple metabolic networks, provide an ideal model for investigating universal principles of metabolism and the evolution and function of metabolic networks.

The question driving this research is how different bacterial species accomplish similar metabolic tasks, and to what extent these tasks are conserved. By focusing on specific metabolic pathways, such as carbon metabolism and biosynthesis of amino acids, this research aims to build and compart network models for ten species of bacteria. Through these models, we will explore their structure and connectivity to investigate how network properties and resource allotment vary across species.

In addition to addressing gaps in our understanding of the complexity of these systems, this research has broader implications for the field of synthetic biology. Uncovering principles that govern metabolism could lead to advances in this field, which aims to engineer biological systems for specific and varied purposes.

RELATED WORK

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.

DATASET

The dataset for this research was sourced from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database which is a well-establish and commonly used resource in bioinformatics research (https://www.kegg.jp/kegg/). KEGG provides comprehensive information for a large number of organisms, including information on networks, genes, genomes, reactions, enzymes, compounds, and more. This research focuses on the datasets related to metabolic networks, reactions, and enzymes for ten bacterial species.

- Escherichia coli (‘eco’)
- Bacillus subtilis (‘bsu’)
- Staphylococcus aureus (‘sau’)
- Mycobacterium tuberculosis (‘mtu’)
- Pseudomonas aeruginosa (‘pae’)
- Salmonella enterica (‘sen’)
- Helicobacter pylori (‘hpy’)
- Lactobacillus acidophilus (‘lba’)
- Streptococcus pneumoniae (‘spn’)
- Corynebacterium diphtheriae (‘cdi’)

These species of bacteria are amongst the most well-studied organisms which reduces the likelihood of missing data and provides the highest likelihood of completeness of the metabolic network data. The KEGG database is structured consistently across species, which facilitates the construction of network models and the analysis of metabolic networks. This consistency, along with the extensive data available for these particular species, ensures that the data is as clean as possible and compatible for cross-species comparisons.

Metabolic networks often contain thousands of reactions and their corresponding metabolites. For this research, we focused on the ten global functional metabolic sub networks available through the KEGG Pathways database as representations for functional modules. This allows us to build and tag metabolic networks based on their function, facilitating the analysis of networks based on high order functionality. This research focused on these subnetworks as they represent function at a higher level and provide us a base upon which further in-depth research can be performed.

- Biosynthesis of secondary metabolites
- Microbial metabolism in diverse environments
- Carbon metabolism
- 2-Oxycarboxylic acid metabolism
- Fatty acid metabolism
- Biosynthesis of amino acids
- Nucleotide metabolism
- Biosynthesis of nucleotide sugars
- Biosynthesis of cofactors
- Degradation of aromatic compounds

Data retrieval from KEGG is facilitated through the KEGG REST API, which allows for selective access to relevant information without the need to download entire datasets for each organism. Custom scripts were developed to systematically retrieve and build metabolic networks for each selected organism. These scripts ensure that the entirety of the information pertaining to the subnetworks for each organism are retrieved by leveraging the interconnected nature of the KEGG data.

MAIN TECHNIQUES

The analysis performed in this research required construction of networks for each organism and each sub network. This construction consisted scripts that accessed the REST API of the KEGG database for the organisms and pathways described in the previous section. This preparation and the consistent and robust nature of the KEGG database limited the need for many data mining techniques. Analysis of mined data showed no missing values for any of the organisms and subnetworks.

Preparation of the data by including subnetwork and organism for each entry allowed for tagging and classification based on our predetermined sections (i.e. organism and subnetwork). This can be seen in figure 2 where the subnetworks for the organism Escherichia coli are distinguished by color.

Each entry of the data obtained from the KEGG database had the following categories that were used in the construction of the metabolic networks.

Organism

Pathway

Reactions

Substrates

Products

Construction of the network structure was done using the python package NetworkX. NetworkX is a python library designed for the creation, manipulation, and study of complex networks of nodes and edges (graphs). A directed bipartite network was created with two classes of nodes: reactions and metabolites (substrates and products). Directed edges went from substrates to reactions, or from reactions to products. Metabolic networks are networks where molecules (called metabolites) go through a series of reactions to either provide a needed molecule for a certain function, or to provide energy to be used elsewhere in the cell. Due to the series like nature of metabolic networks, products in one reaction are substrates in another leading to an interconnected network that can be modeled as described.

Once the network model was constructed for each organism, it was filtered for the largest connected cluster. Due to limitations in the ability to fully map out metabolic networks and to the variable nature of functional modules across species, portions of the metabolic sub networks were unconnected to the rest of the graph. Many network metrics have the constraint that the graph must be fully connected (every node must attach to some other node in the graph), so filtering for the largest connected cluster removes smaller, unconnected portions of the networks and allows for the calculation of these metrics. Figure 1 below shows the unfiltered metabolic network (a) compared to the filtered network (b).

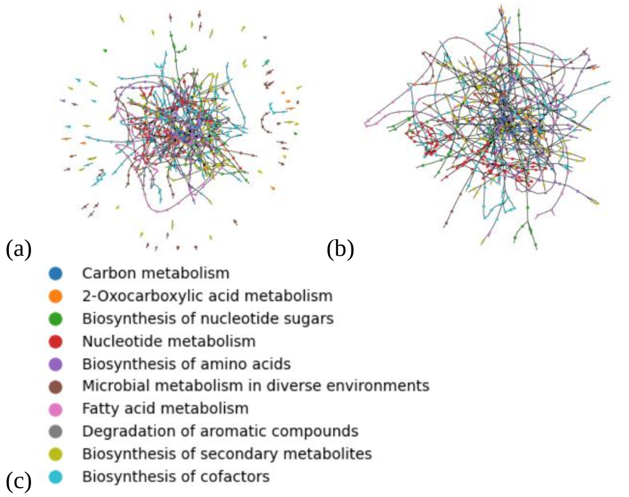
(a)
(b)
(c)
● Carbon metabolism
● 2-Oxocarboxylic acid metabolism
● Biosynthesis of nucleotide sugars
● Nucleotide metabolism
● Biosynthesis of amino acids
● Microbial metabolism in diverse environments
● Fatty acid metabolism
● Degradation of aromatic compounds
● Biosynthesis of secondary metabolites
● Biosynthesis of cofactors

Figure 1: visualization of the metabolic network (containing only the functional modules described previously) for Escherichia coli. (a) unfiltered network. (b) filtered network. (c) node colors of the different functional modules.

Networks were also constructed for individual functional subnetworks using the same method. Data was grouped by organism and by functional pathway and then filtered to remove unconnected parts allowing for the calculation of different network metrics. Figure 2 to the right shows the subnetwork graph visualization of the Nucleotide Metabolism subnetwork for different species. The variability in the structure and folding of the graph is due to the inherent variability between these species, even for the same metabolic function.

KEY RESULTS

This research focused on the exploratory construction and analysis of bacterial metabolic networks, with the goal of assessing the network structure of the models to provide insights into biological function. The process began with the development of scripts to retrieve and organize data from the KEGG database, followed by construction of metabolic networks using NetworkX. These networks were then analyzed across multiple bacterial species to identify patterns of conservation and divergence in the network structure of different subnetworks present within each species.

To analyze network structure, the following network metrics were used;

- *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
- *Number of Nodes*: used to measure the distribution of resources (genes corresponding to enzymes that facilitate each reaction) each organism allots to each functional subnetwork.

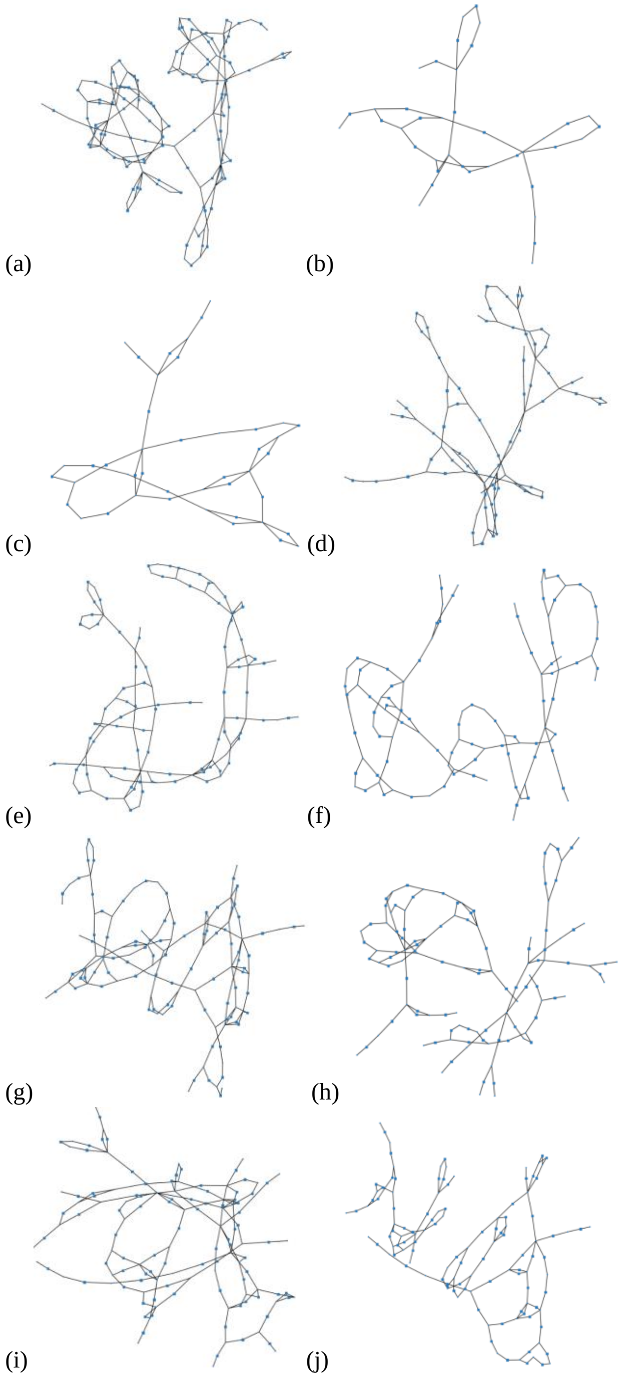

Figure 2: Network visualization for Nucleotide Metabolism across species (the following codes are described in the dataset section). (a) eco (b) hpy (c) lba (d) mtu (e) pae (f) sau (g) sen (h) spn (i) bsu (j) cdi

Figure 3 below shows the average value across all of the ten subnetworks for each of the metrics above where each bar represents one of the ten organisms studied. Comparing the Average Path Length chart with the Density chart in figure 3 seems to show an inverse relationship between those two metrics. Organisms with a low average path length (hpy, lba) have the highest density. This is not unexpected, as density corresponds to the proportion of potential connections that are actually present whereas average path length corresponds to the average of shortest path distances between all possible pairs in the network. If a network has a high density, then there are (proportionally) more connections between nodes, and if there are more connections between nodes, there are more potential connections from which a shortest path between points can be drawn. Both of these metrics have an influence on how something can travel through a network (in this case this is most often measured by carbon molecules as most metabolites are carbon based or at least contain carbon), so in that context, an inverse relationship would indicate that longer metabolic networks make up for the efficiency loss of higher distances from one point to another, by increasing the density and thus robustness of the network so that any disruption farther away from a node doesn't have as many far down stream effects.

Comparing the Density chart to the Number of Nodes chart shows an exact inverse relationship. The more nodes present, the lower proportion of possible connections actually present in the graph. This makes sense as the number of possible connections (edges) in a graph grows exponentially with the number of neurons. There doesn't seem to be any correlation between the Number of Nodes and the Average Path Length.

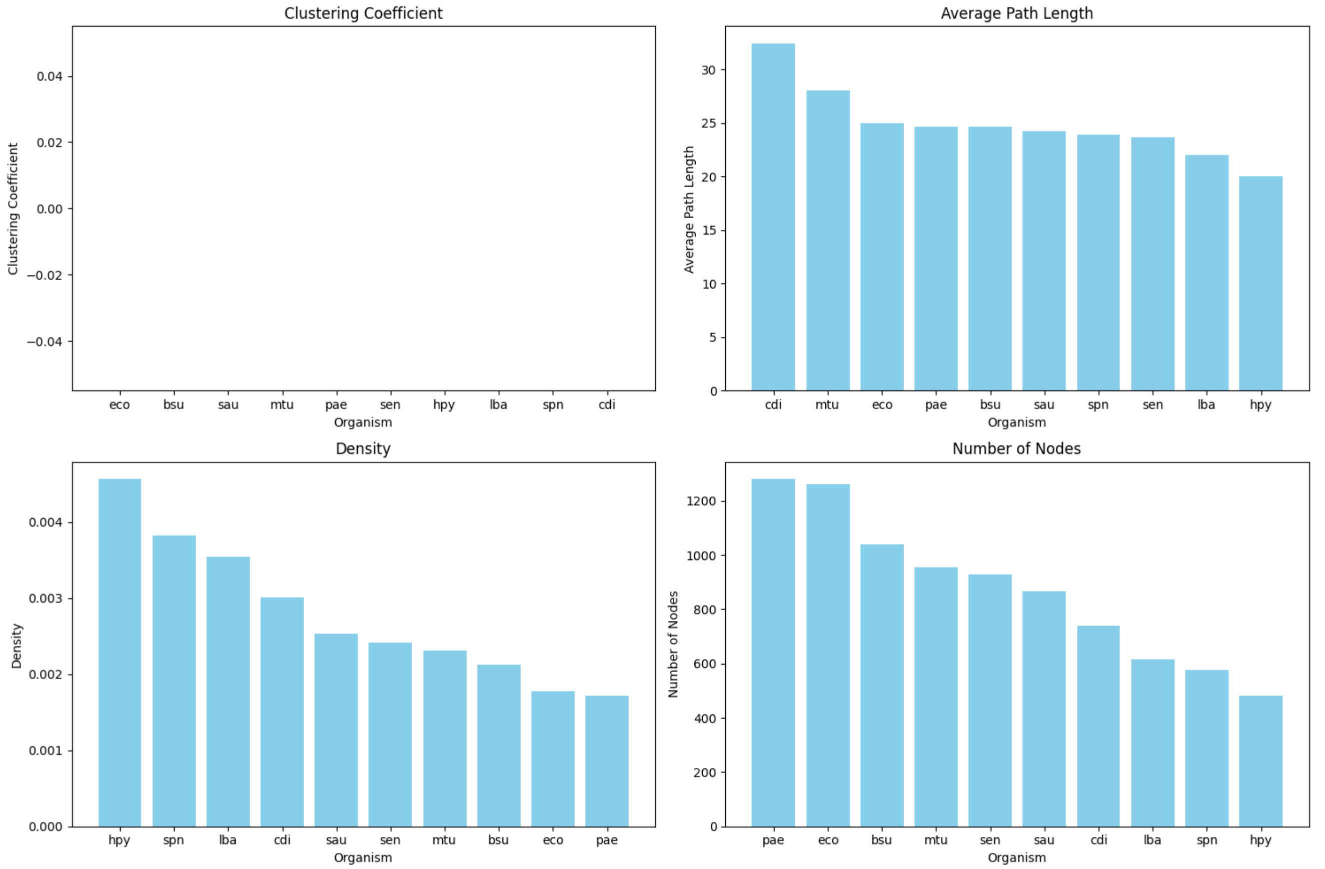

Figure 3: Average values for each network metric across all subnetworks. Each bar in the charts represent each of the ten organisms.

The clustering coefficient is a measure of the number of triangles (cycles of 3) in the graph. The nature of bipartite graphs is that nodes of one type can only have edges to nodes of another type (reactions and metabolites in our case), so the smallest possible cycle is four (metabolite1 -> reactionA -> metabolite2 -> reaction -> metabolite1). Thus the 0's across the board for the clustering coefficient are due to the inherent nature of the graph representation chose. It is possible to collapse the graph down to one of the node types which would allow for calculation of clustering coefficients, but there wasn't sufficient time to carry this out.

Figure 4 below shows the networks metrics, this time broken down by organism and subnetwork. From this graph, we can see that there is much more variation across these metrics between subnetworks than there is between species, indicating that there is more conservation of network properties in functional subnetworks than there is across species overall metabolic networks (this could potentially be due to the smaller size of these subnetworks as smaller sample sizes inherently have more deviation).

Looking at the Number of Nodes by Pathway chart in figure 4, we can see that the amount of metabolic resources (measured as reactions corresponding to genes encoding enzymes) are fairly consistent across species with the most resources being allotted to

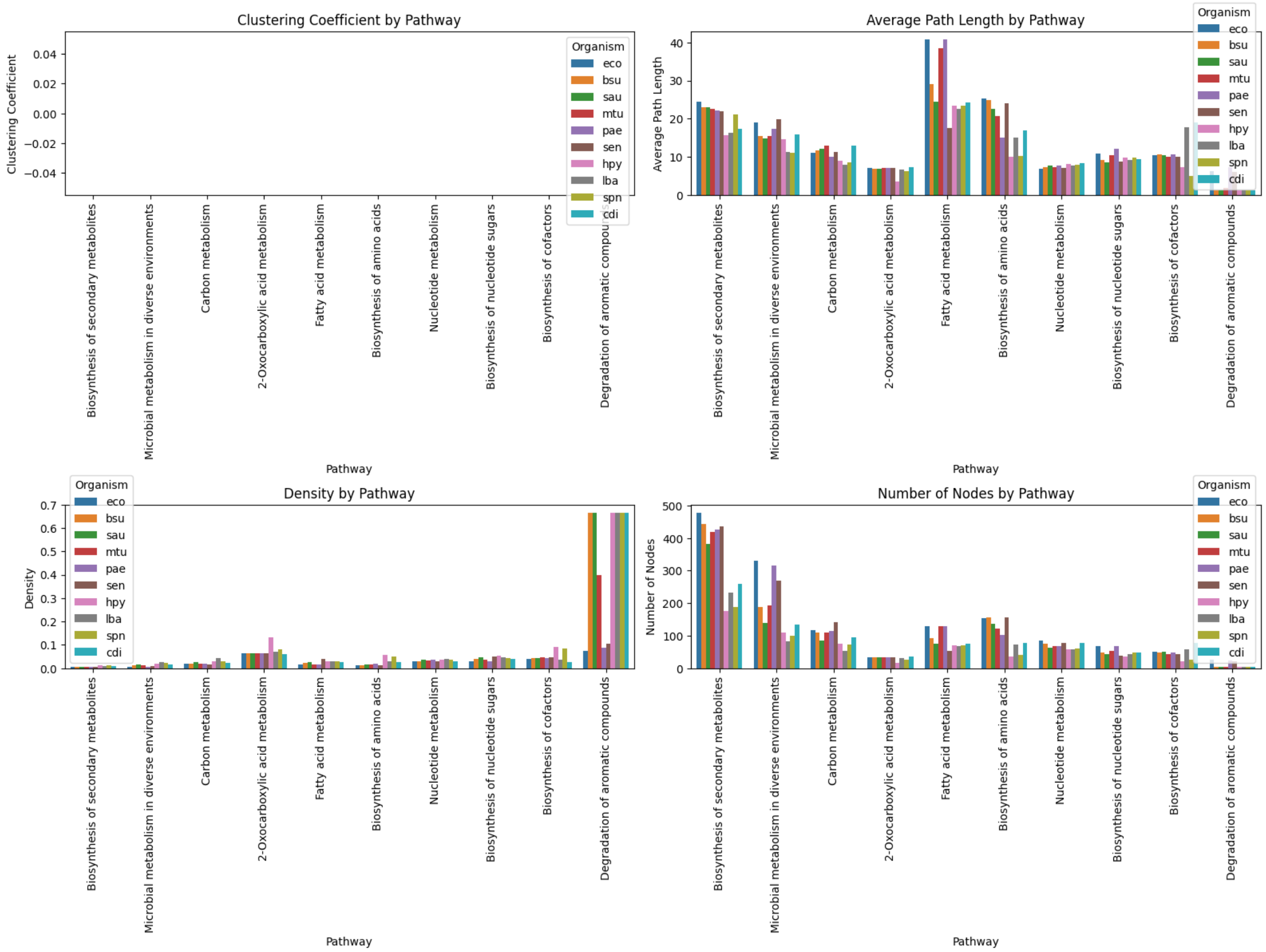

Figure 4: Network metrics broken up by functional subnetwork (x axis) and by organism (color). X axis labels are (in order): Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments, Carbon metabolism, 2-Oxycarboxylic acid metabolism, Fatty acid metabolism, Biosynthesis of amino acids, Nucleotide metabolism, Biosynthesis of nucleotide sugars, Biosynthesis of cofactors, Degradation of aromatic compounds.

Biosynthesis of secondary metabolites and metabolism in diverse environments, and the least amount of resources allotted to 2-Oxycarboxylic acid metabolism and Degradation of aromatic compounds.

Secondary metabolites are compounds produced by organisms that are not directly involved in the normal growth, development, or reproduction of that organism. While those are all biological imperatives that require significant investment, secondary metabolites could potentially be associated with complexity as they allow for more exploration of metabolic space that could contribute to an organisms overall ability to carry out those biological imperatives. Thus it would make sense that organisms would devote a fair amount of resources to complexity as the potential to explore more of any space is generally an evolutionary advantage. 2-Oxocarboxylic acids are a vital metabolites for synthesis and degradation of amino acids (building blocks of proteins) as well as for the TCA cycle which generates most of the cells energy, so the common low allotment of resources to this subnetwork is somewhat surprising. The high allotment of resources to metabolism in diverse environments is somewhat surprising and indicates favorability of robustness in multiple environments. Organisms face the potential of changing environments constantly, so evolution favors those organisms that are capable of surviving changes in their environment.

In comparing the Number of Nodes chart with the Density chart, we see the same exact inverse relationship as we did in figure 3 due to the same effect of number of possible edges scaling exponentially with the number of nodes in a graph.

In the Average Path Length chart, we can see that this value is relatively similar for each organism, but differs significantly by functional subnetwork. As mentioned previously, average path length is a measure of shortest path length between all possible nodes, and corresponds to how efficiently flow can move through the network. This value was generally highest for fatty acid synthesis and biosynthesis of amino acids. Both of these compounds play important roles in many parts of a cell, and so this increased focus on efficient flow through these networks indicates that efficiency is prioritized for functions that have many roles.

Overall, this research contributes to our understanding of bacterial metabolic networks by highlighting how organisms organize and prioritize metabolic resources. Although not much variability between species was seen, the network metrics by functional group provided insights into how the network structure of the metabolism corresponds to functional roles.

APPLICATIONS

The dream of synthetic biology is to engineer biological systems to carry out specific functions. In order to realize this though, we have to understand more about the complexity of biological systems. This complexity is due to the interconnected nature of biological systems at every level, and without understanding it more deeply, it is unlikely we will ever reach the engineering level that we have with simpler systems. Current limitations on technologies in this field arise in part because biology developed with a focus on the whole rather than the single points engineering is commonly focused on. Robustness and resiliency are properties of biology that limit the impact that any single point has on the system as a whole. If we can understand more about how the interconnectedness of single points in a system affect and interact with the rest of the system, then we could begin to have more control over engineering such systems. This research provides a foundation for further research to elucidate more of how network structure plays a role in and impacts the functionality of subnetworks in the metabolism.