**BENG 213 - Metabolic Promiscuity in *E.coli* - Checkpoint 1**

Jason Hyun, Kevin Glass

**Literature Review**

E. coli is a commonly studied gram negative bacterium commonly used in lab experiments, and has had its metabolic network reconstructed repeatedly [1]. The model of E. coli we used for this project was iML1515, which features 1,515 open reading frames, 2719 reactions with 1,192 unique metabolites. The model is also linked to 1,515 protein structures [2], which allows us to incorporate and analyze promiscuity as being related to variety of data types provided by the model, such as protein structural information, metabolite structural information, gene activity, network connectivity, and enzyme family. Enzyme promiscuity, as defined by enzymes catalyzing multiple different reactions for which they are not normally recognized, can be quantized using multiple different approaches chemical [3], however for our purposes, we are defining promiscuity to be one gene set catalyzing more than one reaction. Collectively, the metabolism of promiscuous enzymes beyond their normal recognized catalytic function can be referred to as a cell’s “Underground metabolism” [5]. Previous models have been generated of E. coli’s underground metabolism, incorporating existing enzyme chemical information to an existing genome scale model of E. coli from Brenda [4]. Enzyme promiscuity data for this reaction network was generated to study the side reactions present in the network and to better understand the shadow metabolism, and thus how E. coli might respond to different conditions.

Enzyme promiscuity data has been previously incorporated into genome scale models of E. coli [4], and has been used to try and understand the effects of the underground metabolism on network dynamics. Furthermore, previous studies have also classified enzymes within E. coli as generalists (catalyzing more than one reaction) versus specialists (only catalyzing one reaction) and have analyzed various network properties along this spectrum of specificity [6]. Furthermore, metabolic networks have been analyzed with in vivo experimentation to better understand promiscuity within a network, and determine previously unknown underground activity [7]. These analyses have either incorporated enzyme promiscuity data into known biochemical reaction networks, analyzed promiscuity from a network perspective, or have tried to experimentally determine promiscuity based on in-vivo knockouts. Previous analysis have yielded that promiscuous reactions can be incorporated into existing reaction networks, and can provide advantages for growth under specific conditions, can be predicted using a combination of gene knockouts and the already established metabolic network, and can provide resistance to perturbations in a cells requirement and may require less complex regulation. Thus, enzyme promiscuity in a network has been studied in the context of experimental gene knockouts, metabolic network structure, metabolite similarity, protein structural information, and fluxomic information.

Open areas of research are whether enzyme promiscuity can be predicted using supervised learning methods by utilizing structural, fluxomic, metabolic, gene knockout information, and whether existing reactions can provide a framework for discovering unknown promiscuous reactions. Furthermore, can we use a combination of these factors to determine possible areas within a metabolic network that promiscuity might exist? Also, as generalist and specialist enzymes have been hypothesized to evolve from duplicates of a given enzyme, can enzyme promiscuity be understood in the context of relatedness to a specialized, high flux enzyme, or pathway position? Furthermore, possible applications of group contribution for calculating metabolite similarity, and ergo likely substrate promiscuity is also an area of potential exploration.

**Data Collection and Visualization**

*Defining enzyme and promiscuity directly from iML1515*

Our analysis to date is based on the iML1515 reconstruction of *E. coli* [2]. We first defined an enzyme as a set of gene(s) able to independently catalyze at least one reaction in iML1515. This was accomplished by iterating through each reaction in iML1515, splitting the boolean GPRs at each ‘OR’, treating each gene (or group of genes if ‘AND’s are present) as an enzyme, and assigning the reaction to the enzyme. Multiple reactions were assigned to an enzyme if its gene group was encountered more than once during this process. This yielded 1242 enzymes, of which 974 were not related to transport. Each enzyme was also verified to be consistent with the GPR of every reaction assigned to that enzyme. Finally, Uniprot accession IDs for each gene was extracted from iML1515-GEMPro and assigned to the appropriate enzyme(s); while not analyzed at this checkpoint, these annotations may be used in the future to integrate protein structure data.

We defined promiscuity as the number of reactions assigned to an enzyme (i.e. the number reactions the enzyme can catalyze independently). **Figure 1** shows a histogram of enzyme promiscuities for non-transport enzymes, ranging from 1 to 41 reactions. About two-thirds of the enzymes were specific, and most promiscuous enzymes catalyzed 5 or fewer enzymes.

*Constructing a substrate-based similarly/distance matrix between enzymes*

In order to quantify the relationships between enzymes encoded in iML1515, we examined the substrates associated with each enzyme. For a given enzyme, we defined its substrate set as the union of all substrates across all reactions associated with the enzyme. Reaction products were also included for reversible reactions, and H2O and H+ were excluded. A similarity matrix was constructed for all pairs of enzymes using the Jaccard index between their substrate sets. Similarly, a distance matrix was constructed as 1 - Jaccard index.

The Jaccard index was used as a simple exploratory metric to compare enzymes, based on the hypothesis that enzymes that share substrates or reactions may have similar structures / active sites and possibly similarly promiscuity properties. Future work may expand on this step to integrate actual structural similarity using BIGG IDs to look up metabolite structures and reaction EC numbers, and/or protein sequence or structure similarity from the Uniprot annotations.

*Visualization of Jaccard distance matrix*

We applied three approaches to visualize the substrate-based Jaccard distance matrix: PCA, tSNE, and heatmap with hierarchical clustering. For the PCA plot **(Figure 2)**, data appeared to be structured along three directions when plotted along the first two components; however, the structure appeared independent of enzyme promiscuity, and the first two components captured very little total variance (27.6% and 6.9%). Curiously, when generating a similar PCA plot for a z-score normalized distance matrix, even less variance was captured by the first two components (11.9%, and 5.0%) and again any visible structure appeared independent of enzyme promiscuity. This is likely due to distance matrix being not normally distributed and bounded between [0,1] from the definition of Jaccard index, making z-score normalization not appropriate for this data.

Similarly for the tSNE plot **(Figure 3)**, the data formed a single large cluster along with several distant, smaller clusters, but most clusters did not segregate on the basis of enzyme promiscuity. This structure was observed across multiple tSNE visualizations.

For the heatmap visualization, we first clustered the enzymes using three types of hierarchical/agglomerative clustering methods that are compatible with non-Euclidean distances: Nearest Point (single), UPGMA (average), and WPGMA (weighted). **Figure 4** shows dendrograms and heatmaps for these approaches. Clusters or heatmap blocks (i.e. groups of enzymes that share many substrates among themselves) generated by Nearest Point were much different from those derived by UPGMA or WPGMA, which had similar results. UPGMA/WPGMA also appeared to have identified about 5 relatively large clusters, as well as a number of smaller clusters. An interpretation of clusters is in the following section.

**Unsupervised Learning**

*Examining cluster consistency across different clustering methods*

In addition to the hierarchical clustering approaches used for heatmap visualization, we also examined DBSCAN as an alternative approach. These methods were selected as they are able to handle non-Euclidean distances (i.e. Jaccard distances) and uneven cluster sizes, particularly important to preserve smaller clusters in the presence of larger clusters that may form simply due to shared cofactors.

We started with DBSCAN to better understand the appropriate number of clusters to extract from hierarchical clustering. **Table 1** shows the number of clusters (not including the “unassigned” cluster) inferred by DBSCAN with different neighborhood sizes and minimum cluster sizes. We found that 13 clusters were observed for multiple settings, (ε = 0.5, min\_size = 6 or 7; ε = 0.7, min\_size = 5; ε = 0.9, min\_size = 3) and that below 9 clusters, the total number of clusters was unstable with respect to DBSCAN parameters (i.e. for epsilon = 0.7, total clusters observed would fluctuate up and down between 9 and 4 as the minimum cluster size increased from 6 to 12). Additionally, it is possible to infer 10-13 clusters by visual inspection of the tSNE plot from **Figure 3**. From this, we adjusted the cut threshold of each hierarchical clustering method to produce 13 clusters. We compared clusters generated using six methods:

1. Hierarchical Clustering, Nearest-Point, 13 clusters
2. Hierarchical Clustering, UPGMA, 13 clusters
3. Hierarchical Clustering, WPGMA, 13 clusters
4. DBSCAN, ε = 0.5, minimum cluster size = 6 (13 clusters + 769 unassigned)
5. DBSCAN, ε = 0.7, minimum cluster size = 5 (13 clusters + 386 unassigned)
6. DBSCAN, ε = 0.9, minimum cluster size = 3 (13 clusters + 132 unassigned)

To evaluate how consistent clusters were across these approach, we computed the Rand Index between each pair of clusterings (i.e. the fraction of pairs of elements that are either in the same cluster or in differents clusters for both clusterings). **Figure 5** shows a heatmap of the pairwise Rand Indexes. As previously observed for Hierarchical Clustering, we found that UPGMA and WPGMA produced very similar clusterings that were distinct from the Nearest-Point clustering. Additionally, we found that the DBSCAN clustering with large neighborhoods (ε = 0.9) was similar to that of UPGMA/WPGMA hierarchical clustering, while DBSCAN with smaller neighborhoods was quite different; however, this may be due to the large number of unassigned data points when using smaller neighborhoods in which the Rand Index would inflate how different two clusterings are. Taking this into account, there appeared to be two distinct clusterings between the methods, one derived from DBSCAN or Hierarchical Clustering with either UPGMA or WPGMA, and another from Hierarchical Clustering with Nearest-Point.

However, upon further analysis, we found that all three Hierarchical Clustering methods had essentially failed to discern significant small clusters, each producing a single very large cluster with all remaining clusters containing just 1-2 points each; in order for for more than one sizable cluster to be observed, more than 100 clusters would have to be extracted from the hierarchy, in which still most clusters would still contain 1-2 points. In contrast, DBSCAN with large neighborhoods (method 6) produced a similar very large cluster, but also several smaller clusters with at least 5 points each. This is significant, as clusters of size 2 may likely be consecutive reversible enzymes in a pathway (i.e. A ↔ B and B ↔ C), while larger clusters can illustrate more complex relationships with regards to shared substrates. This suggests that the Rand Index may not be suitable for comparing clustering in this case as cluster sizes are very uneven and dominated by a single large cluster.

*Evaluating clustering with respect to enzyme promiscuity*

Proceeding with the 13 clusters derived by DBSCAN with large neighborhoods (method 6, shown), we first checked if any of the clusters were enriched for either specific or promiscuous enzymes. In general, the clustering was overall not very predictive of enzyme promiscuity, with a homogeneity of 10.2% and a completeness of 7.9%. However, this may be due to the single large, essentially uninformative cluster generated by all tested clustering approaches; several of the smaller clusters DBSCAN produced were highly enriched for either promiscuous or specific enzymes (**Table 2**).

Among the smaller clusters, notably clusters 1 and 2 were composed of a small number of highly promiscuous enzymes (cluster size << number of unique reactions). Cluster 3, the “largest” small cluster was composed a set of redundant, promiscuous enzymes involved in Murein Biosynthesis and Recycling. Overall, we found that many of these smaller clusters are enriched for either promiscuous or specific enzymes, but for the most part produced a clustering along subsystem annotations.

Finally, for evaluation of clustering independent of promiscuity labels, we applied the same PCA and heatmap visualization from before to the DBSCAN clusters (**Figure 6**). DBSCAN with large neighborhoods seemed to have failed to reproduce many of the loosely related blocks that was visible from the hierarchically clustered heatmaps (**Figure 4**), though it did manage to isolate smaller, more tightly connected blocks (darker blocks in bottom right of heatmap). Additionally, the small clusters were not well separated based on the PCA plot, with almost all of the remaining variance occuring in the largest, uninformative cluster. This is consistent with the Silhouette coefficient (ratio of mean intra-cluster distance to mean nearest-cluster distance) being just 3.8%. It is possible that the nature of this dataset with non-Euclidean, bounded distances and very uneven cluster size distribution is to blame for the poor clustering performance

**Conclusions**

Overall, one of the earliest obstacles in tackling the question of enzyme promiscuity from the iML1515 reconstruction was the problem of extracting numerical data from an unweighted network with qualitative/categorical annotations. As an exploratory approach, we inferred enzymatic units directly from reaction GPRs, and assigned to each enzyme all reactions it could independently catalyze. We then computed a pairwise distance matrix between enzymes based on the Jaccard index applied to sets of substrates each enzyme could act on.

This method provided us a starting point to which we could apply common data visualization and unsupervised learning methods. However, the use of non-Euclidean distances and a highly uneven cluster size distribution present in the data complicated these analyses; particularly, the first two PCA components captured very little variance, multiple types of hierarchical clustering produced a single large cluster and several clusters of size 1-2 (unless >100 clusters were extracted), and DBSCAN had poor separation and failed to identify much of the structure visible from hierarchical clustering. While some DBSCAN clusters were enriched for promiscuous enzymes, the clustering was not much more informative than examining reaction subsystems.

Ultimately, examining just the Jaccard distances between enzymes’ known substrates does not appear to be very informative with regards to promiscuity. While revisiting this analysis with a much larger number of clusters (as was visible from hierarchical clustering) may expose additional, smaller patterns between substrate and promiscuity, we do not think it is likely that just this data will help identify novel candidates for promiscuous enzymes. Moving forward, additional network and structural information such as distances in the metabolic network, simulated fluxes, metabolite structures, and protein sequences/structures may be integrated to the analysis, making use of the BIGG and Uniprot IDs extracted in this analysis.

**References**

All analysis was implemented in Python3.6. Extraction and definitions of enzymes from iML1515 was implemented in part with COBRApy. Hierarchical clustering was implemented through SciPy. PCA, tSNE, and DBSCAN were implemented through scikit-learn (sklearn).

[1] McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol*. 9:661. doi: 10.1038/msb.2013.18.

[2] Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BØ (2017). iML1515, a knowledgebase that computes Escherichia coli traits. *Nat Biotechnol*. 35(10):904-908. doi: 10.1038/nbt.3956.

[3] Khersonsky O, Tawfik DS (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem.*79:471-505. doi: 10.1146/annurev-biochem-030409-143718.

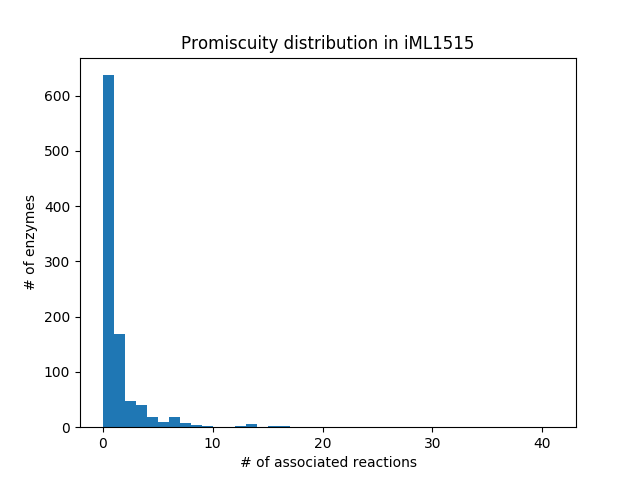
[4] Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, Lázár V, Spohn R, Csörgő B, Wagner A, Ruppin E, Pál C, Papp B (2014). Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci USA*. 111(32):11762-7. doi: 10.1073/pnas.1406102111.

[5] D’Ari R, Casadesus J (1998). Underground metabolism. *BioEssays*. 20:181–186

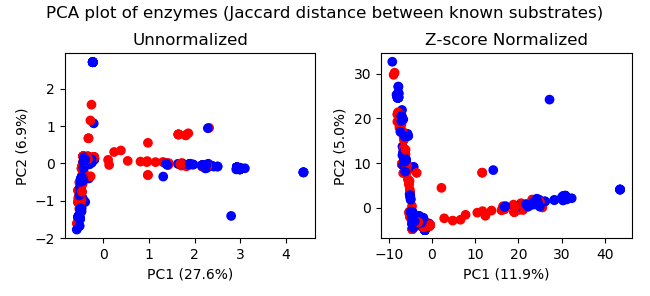
[6] Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, Kim D, Palsson BØ (2013). Network context and selection in the evolution to enzyme specificity. *Science*. 337(6098): 1101–1104. doi: 10.1126/science.1216861.

[7] Guzmán GI, Utrilla J, Nurk S, Brunk E, Monk JM, Ebrahim A, Palsson BØ, Feist AM (2015). Model-driven discovery of underground metabolic functions in Escherichia coli. *Proc Natl Acad Sci USA*. 112(3):929-34. doi: 10.1073/pnas.1414218112.

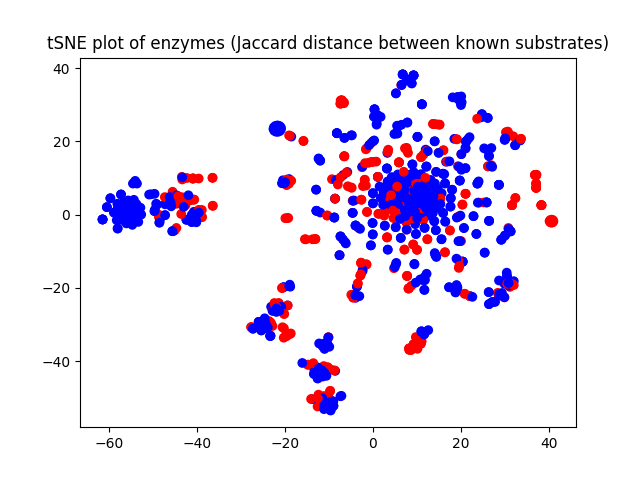
**Figures and Tables**



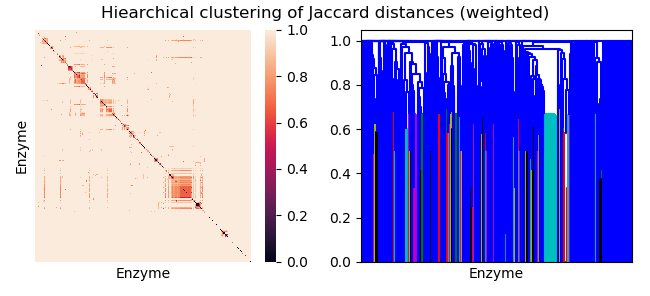
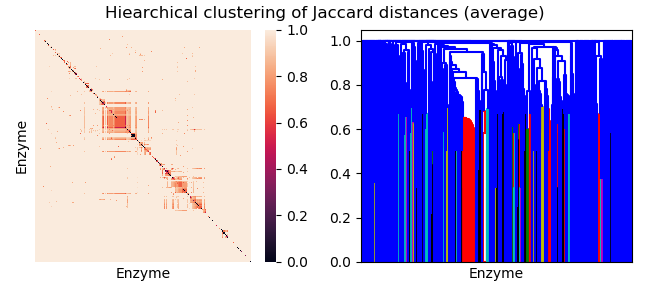
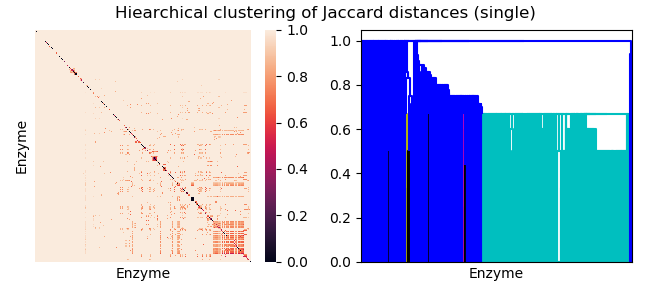
**Figure 1:** Histogram of promiscuity for 974 non-transport enzymes extracted from iML1515.



**Figure 2:** PCA (with and without z-score normalization) plot of substrate-based Jaccard distance matrix for 974 non-transport enzymes extracted from iML1515. Promiscuous enzymes (promiscuity > 1) are shown in red, specific enzymes are shown in blue.

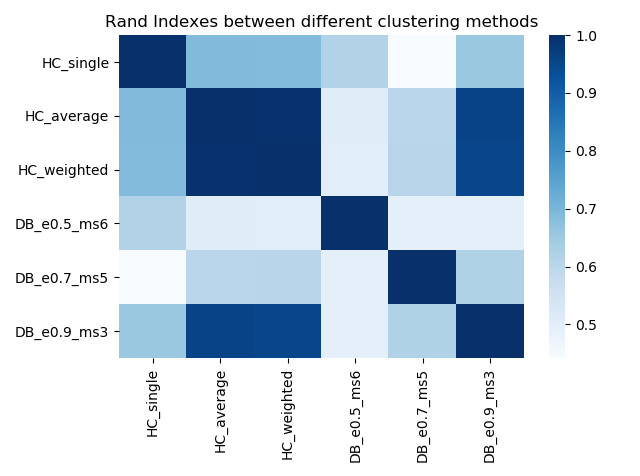


**Figure 3:** tSNE visualization of substrate-based Jaccard distance matrix for 974 non-transport enzymes extracted from iML1515. Promiscuous enzymes (promiscuity > 1) are shown in red, specific enzymes are shown in blue.



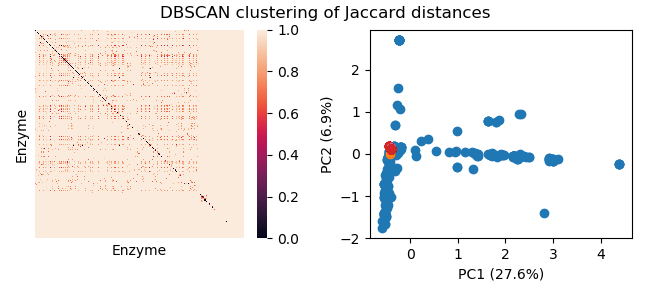
**Figure 4:** Heatmap and dendrogram of substrate-based Jaccard distances after hierarchical clustering of 974 non-transport enzymes extracted from iML1515. Nearest Point, UPGMA, and WPGMA was used for clustering.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1:** Number of enzyme clusters inferred by DBSCAN from Jaccard distance matrix using different neighborhood sizes (ε) and minimum cluster sizes. | | | | | | | | | | | |
| **ε = 0.5** | | | | | | | | | | | |
| Min. Cluster Size: | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| # Clusters: | 135 | 63 | 35 | 21 | 13 | 13 | 7 | 5 | 4 | 4 | 3 |
| **ε = 0.7** | | | | | | | | | | | |
| Min. Cluster Size: | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| # Clusters: | 71 | 34 | 25 | 13 | 7 | 9 | 8 | 5 | 5 | 6 | 4 |
| **ε = 0.9** | | | | | | | | | | | |
| Min. Cluster Size: | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| # Clusters | 27 | 13 | 9 | 6 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |

****

**Figure 5:** Rand Indexes between different clustering methods applied to Jaccard distances between the substrates of the 974 non-transport enzymes extracted from iML1515.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2:** Characterization of clusters derived from DBSCAN using ε = 0.9, minimum cluster size = 3. Dominant subsystem is defined as any subsystem that represents at least half of all the unique reactions that are catalyzed by enzymes in a cluster, as annotated in iML1515. The largest two clusters (0 and -1, essentially unassigned enzymes) were not examined. | | | | | |
| Label | Cluster Size | # Promiscuous Enzymes | Percent Promiscuous | # Unique Reactions | Dominant Subsystems |
| 0 | 763 | 261 | 34.21 | - | - |
| 1 | 3 | 3 | 100 | 29 | Cell Envelope Biosynthesis (25/29) |
| 2 | 4 | 4 | 100 | 33 | Nucleotide Salvage Pathway (20/33) |
| 3 | 34 | 34 | 100 | 29 | Murein Biosynthesis (15/29)  Murein Recycling (14/29) |
| 4 | 4 | 3 | 75 | 3 | Alternate Carbon Metabolism (3/3) |
| 5 | 5 | 4 | 80 | 4 | Diverse |
| 6 | 6 | 2 | 33.33 | 5 | Nucleotide Salvage Pathway (3/4) |
| 7 | 5 | 3 | 60 | 6 | Lipopolysaccharide Biosynthesis / Recycling (4/6) |
| 8 | 3 | 1 | 33.33 | 2 | Diverse |
| 9 | 5 | 1 | 20 | 6 | Cofactor and Prosthetic Group Biosynthesis (6/6) |
| 10 | 3 | 0 | 0 | 3 | Diverse |
| 11 | 3 | 0 | 0 | 3 | Cell Envelope Biosynthesis (3/3) |
| 12 | 4 | 0 | 0 | 4 | Lipopolysaccharide Biosynthesis / Recycling (4/4) |
| -1 | 132 | 21 | 15.91 | - | - |



**Figure 6:** Visualization of DBSCAN clustering applied to Jaccard distances. While 13 clusters are present, all except the largest cluster are concentrated in a single spot in the PCA plot.