

## An Evaluation of Different Clustering Methods and Distance Measures Used for Grouping Metabolic Pathways

Sarah M. Kim<sup>\*</sup>, Matthew I. Peña<sup>†</sup>, Mark Moll<sup>\*</sup>, George Giannakopoulos<sup>‡</sup>,  
George N. Bennett<sup>†</sup>, Lydia E. Kavraki<sup>\*</sup>

<sup>\*</sup>Department of Computer Science, Rice University

<sup>†</sup>Department of BioSciences, Rice University  
Houston, TX, 77005, USA

<sup>‡</sup>Software and Knowledge Engineering Lab,

ITT NCSR Demokritos  
Aghia Paraskevi 15310, Greece

### Abstract

Large-scale annotated metabolic databases, such as KEGG and MetaCyc, provide a wealth of information to researchers designing novel biosynthetic pathways. However, many metabolic pathfinding tools that assist in identifying possible solution pathways fail to facilitate the grouping and interpretation of these pathway results. Clustering possible solution pathways can help users of pathfinding tools quickly identify major patterns and unique pathways without having to sift through individual results one by one.

In this paper, we assess the ability of three separate clustering methods (hierarchical,  $k$ -means, and  $k$ -medoids) along with three pair-wise distance measures (Levenshtein, Jaccard, and  $n$ -gram) to expertly group lysine, isoleucine, and 3-hydroxypropanoic acid (3-HP) biosynthesis pathways. The quality of the resulting clusters were quantitatively evaluated against expected pathway groupings taken from the literature.

Hierarchical clustering and Levenshtein distance seemed to best match external pathway labels across the three biosynthesis pathways. The lysine biosynthesis pathways, which had the most distinct separation of pathways, had better quality clusters than isoleucine and 3-HP, suggesting that grouping pathways with more complex underlying topologies may require more tailored clustering methods.

## 1 Introduction

The size and number of available metabolic databases such as KEGG [15, 16], MetaCyc [3], BRENDA [26], and Reactome [5] continue to rapidly expand as new discoveries about metabolism are made and old studies are mined for useful data. This wealth of metabolic information significantly improves the chances of discovering novel and valuable heterogeneous pathways. However, manually searching for these pathways is a tedious task, providing the motivation for automating metabolic pathwayfinding. Several algorithms and tools have been developed which utilize available metabolic data to find novel heterogeneous metabolic pathways. Using

graph-based models is one common approach for identifying metabolic pathways and is capable of generating tens of thousands of different results (eg. [1, 4, 7, 22, 24, 25]). A weakness of existing pathfinding methods is that they do not facilitate interpretation of the resulting pathways, leaving the user to sift through overwhelming amounts of information. Current pathfinding software could be improved by providing functionality to the end user by allowing for (1) grouping pathways based on similarity so that results can be more easily navigated and filtered, (2) identifying unique pathways that differ from existing, endogenous pathways, and (3) more effectively displaying a large number of results.

In our previous work, resulting pathways were ranked by length, and exploration of the solutions was limited to displaying pathways one at a time [10]. This visual approach scales poorly when the solutions number in the thousands. Even displaying the top hundred results simultaneously, as done by MetaCyc, may cause the user to overlook unique pathways or patterns that can only be observed across a thousand pathways. Metabolic engineers will ultimately want to narrow down the list of candidate pathways *in silico* before the pathways are designed *in vivo*. Here, we provide a baseline analysis of how well standard clustering techniques can group results from pathway finding searches based on internal and external evaluation measures.

Previous studies have compared and analyzed the clustering of other biological data, even as recently as 2015 by Wiwie et al. [28]. However, the specific case of clustering metabolic pathways has not been examined and is worth further study, given that an understanding of metabolic pathways is essential to many biochemistry-related fields.

## 2 Resources and Methods

Metabolic pathway results were clustered using three different distance measures and three clustering methods implemented in existing R packages [23]. These clusters were then evaluated using internal and external clustering measures. The metabolic pathways, distance measures, clustering meth-

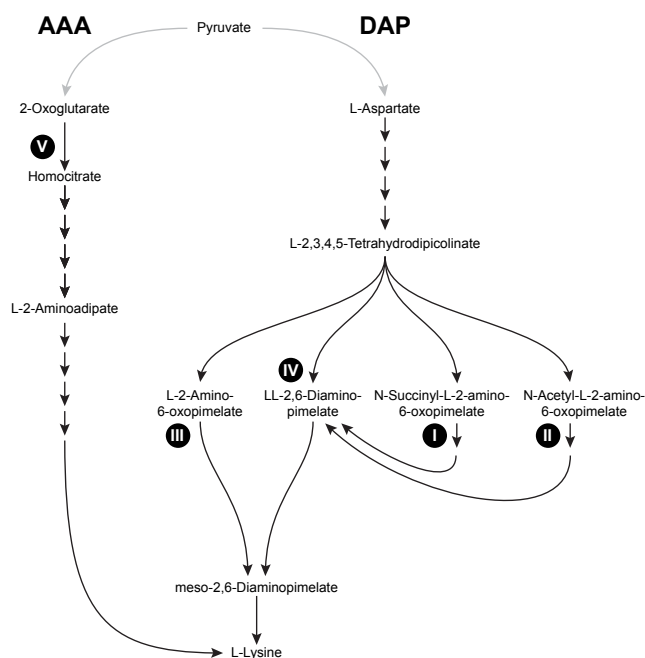


Figure 1: The five expected pathway groups used for external evaluation of lysine clusters. Four of the five pathway labels (labels 1-4) include DAP, and the fifth pathway label includes AAA.

ods, and internal and external cluster evaluation measures are described in the following sections.

**Metabolic pathways.** A test set of three biosynthesis pathways with diverse network topologies was selected for this study. The biosynthesis of lysine and isoleucine, both key amino acids, is central to metabolism of many organisms. These pathways have several known variations. Lysine biosynthesis occurs by one of two evolutionarily distinct pathways, either via diaminopimelate (DAP) in bacteria, algae, and plants or via  $\alpha$ -aminoadipate (AAA) most prevalent in fungi [3, 16]. Isoleucine, on the other hand, can be produced by multiple pathways starting from different metabolic precursors that almost exclusively feed through 2-oxobutanoate [3]. These variations within pathways sharing common starting and ending compounds make these biosynthesis pathways ideal candidates for examining clustering methods. In contrast to lysine and isoleucine, engineered biosynthetic pathways for 3-hydroxypropionic acid (3-HP) from glucose involve enzymes from multiple organisms. These pathways have been patented by Cargill, an agricultural commodities company. At least seven feasible pathways for 3-HP synthesis with unique intermediates have been evaluated based on ATP utilization, thermodynamic favorability, and redox balance [11, 14]. Since 3-HP is a valuable chemical precursor, its synthesis is of significant interest, and the examination of the engineered pathways provides a contrast to the endogenous pathways of lysine and isoleucine that have differentiated through evolution.

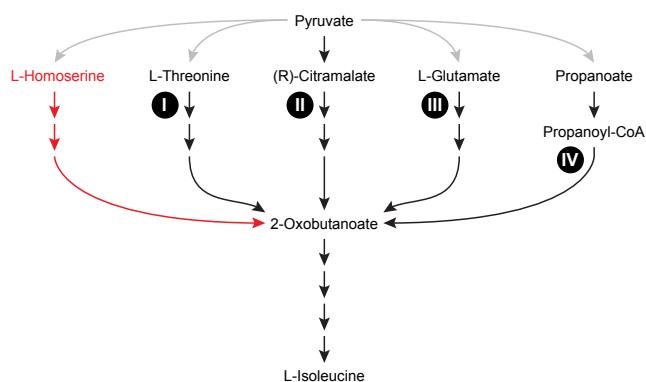


Figure 2: Four expected pathway groups used for external evaluation of the isoleucine clusters. All these pathways include 2-oxobutanoate as an intermediate compound. The fifth expected pathway group does not include 2-oxobutanoate.

For the biosynthetic pathways specific to lysine, isoleucine, and 3-HP, we identified distinguishing pairs of sequential compounds to differentiate pathways that otherwise share common starting and ending points along with a varying number of intermediates. When possible, these distinguishing compounds were selected so as to be consistent with those used by curated metabolic databases (KEGG and MetaCyc) and literature. Lysine biosynthesis can pass through either the aforementioned DAP or AAA pathways, which share no common downstream intermediates. DAP can be further differentiated into four pathways using differences in the series of intermediates between L-2,3,4,5-tetrahydrodipicolinate and diaminopimelate (Figure 1). Using these series of compounds exclusive to each DAP pathway in addition to the primary DAP/AAA dyad provides identifying markers that will serve as labels to which pathfinding solutions can be attributed.

The pathways for isoleucine are not nearly as differentiated as the lysine DAP and AAA pathways because they share a terminal series of compounds common to most pathway variants. The pathways that have been included herein are those described in MetaCyc that pass through 2-oxobutanoate (L-isoleucine biosynthesis pathways I, II, III, and IV). The distinguishing compounds all occur upstream of 2-oxobutanoate. L-threonine, pyruvate, L-glutamate, and propanoate are the starting compounds in the curated MetaCyc pathways (Figure 2). Pyruvate was not selected as a distinguishing compound for this work as it is the starting compound for all pathfinding queries and would therefore not be unique to any subset of clustered pathways. Instead, a compound downstream of pyruvate, (R)-citramalate, has been selected for its uniqueness to the MetaCyc pathway initiated by pyruvate in comparison with the other pathways. Propanoyl-CoA was selected over propanoate because the latter compound was never identified as an intermediate among all pathfinding solutions.

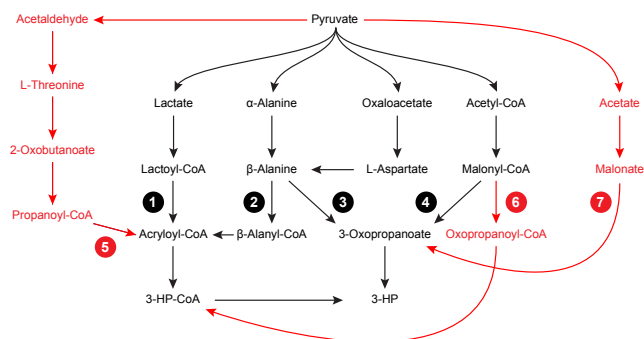


Figure 3: The seven expected pathway groups used for external evaluation of the 3-HP clusters. Pathway group labels 1-4 were taken from Henry et al. [11], while pathway group labels 5-7 were found by the search algorithm and are variations of the known pathways.

In the case of 3-HP, we focused on pathways that originate with pyruvate as it is an intermediate common to all biosynthetic pathways starting from glucose (Figure 3). The substantial overlap of compounds between pathfinding solutions necessitated the use of series of compounds (e.g., lactoyl-CoA to acroyloyl-CoA) rather than individual compounds to compare clustered solutions against distinct pathways described in literature. This resulted in four pairs of labels, to which three additional labels were added to include pairs of compounds that could not be attributed to previously numerated 3-HP biosynthetic pathways from glucose.

**Pathfinding.** Biosynthesis pathway results for converting pyruvate to lysine, isoleucine, and 3-HP were obtained by searching a recent version of the KEGG database (downloaded on Feb. 13, 2015) using the LPAT algorithm developed by Heath et al. [10]. All searches conserved at least one carbon atom from the starting compound to the target compound. The number of pathways found by the algorithm was not limited for any of the searches. A total of 236 pathways, 186 pathways, and 3650 pathways were found for lysine, isoleucine, and 3-HP respectively.

**Computation of pairwise pathway distances.** Pairwise pathway distance measures were computed using Levenshtein distance [27], Jaccard distance [13], and character  $n$ -gram graph distance [9]. These distances were selected to compare different levels of strictness in measuring the similarity between pathways. Levenshtein distance takes the ordering of pathway’s reactions into account, while Jaccard distance disregards ordering and treats the pathway as a set of reactions. The  $n$ -gram graph distance falls between these two approaches by incorporating contextual information about which reactions in the pathways tend to occur together.

To compute Levenshtein distance each pathway was represented as a string of reactions IDs, where each reaction ID was treated as a unique character. The distance between two pathways was quantified as the minimum number of reaction

IDs that needed to be added, removed, or changed in one pathway to make it equivalent to the other [27].

To compute Jaccard distance, each pathway was treated as a set of reaction IDs. Jaccard distance,  $d_J$ , was calculated between each pair of pathways using the following equation:

$$d_J = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Where  $A$  and  $B$  represent the reaction sets of the two compared pathways.

To compute the  $n$ -grams graph distances, pathways were treated as strings of reaction IDs and parsed into  $n$ -grams, given  $n = 3$ . Graphs for each pathway were then constructed using the algorithm described by Giannakopoulos et al. [9]. redDistances were computed by finding the dissimilarity ( $1 -$  normalized value similarity) using a distance window size of three.

**Methods for clustering pathways.** Three methods were used to cluster pathways based on the calculated pairwise distances: agglomerative hierarchical clustering,  $k$ -means clustering, and  $k$ -medoids clustering. These clustering methods are simple and widely used, making them a good baseline for future experiments involving more complex, domain-specific methods.

Agglomerative hierarchical clustering finds clusters by iteratively clumping the two most similar pathways, or groups of pathways, together. The pathways with the smallest pairwise distance are grouped together first. The distance between two groups of pathways was calculated by averaging all pairwise distances between the pathways in the two groups. The results of hierarchical clustering can be presented as a dendrogram, where the number of clusters depends on what level the tree is cut. Hierarchical clustering was selected for being a simple, straightforward clustering measure with easily interpretable results.

Unlike hierarchical clustering,  $k$ -means and  $k$ -medoids clustering requires the number of clusters as input. The  $k$ -means algorithm begins by randomly initializing  $k$  centroids and then assigning each pathway to the closest centroid. Each pathway was treated as a  $n$ -dimensional vector which contained the pathway’s pairwise distances to all other  $n$  pathways. As such, the distance between a pathway and the  $k$ -means centroids, along with the centroids themselves, are not interpretable. Even in  $k$ -medoids, where the centers are chosen from existing pathways, the distances between pathways are not as easily understood as in hierarchical clustering. The  $k$ -means and  $k$ -medoids clustering methods were selected since they are well known and require only the number of clusters and feature matrix as input. The pairwise distance matrix was used as the feature matrix to see if  $k$ -means and  $k$ -medoids would be able to indirectly cluster pathways with non-Euclidean distance measures.

Other clustering methods, such as DBSCAN [6] and Gaussian mixture models [20], were excluded from this study since there was no way in these methods to specify the num-

ber of resulting clusters. In this paper, the number of clusters for each pathway was selected to match the expected number of pathway groupings found in the literature. By fixing the number of clusters, the study’s focus was narrowed to a comparison of the clustering methods and distances measures. Finding a way to determine the optimal number of clusters for a set of pathways without external information is a future topic of interest.

### 3 Evaluation of results

The quality of the clustering results was assessed by (1) internal features based on the distances between pathways in the same and different clusters, and (2) external information based on expected pathway groupings from the literature. The internal evaluation measures were included in this study since clustering pathways is an unsupervised learning problem – groupings for novel pathways found by metabolic pathfinding algorithms often do not exist in the literature, and clustering can only be evaluated using internal measures.

**Internal evaluation.** The Dunn index, silhouette width, adjusted connectivity, and stability of clusters were chosen for the internal evaluation of clusters [2]. Compared to other existing internal evaluation scores, the Dunn index, silhouette width, and adjusted connectivity place more of an emphasis on maximizing the separation distance between clusters and less on the tightness within the clusters. This suits the purposes of this study as there may be many pathway variations within the cluster; however, these variations are not as important as the separation distances between clusters. The first three scoring measures are also robust to different sized clusters. The stability score was included to test that the same clusters will be found even if other pathways are absent from the data set.

- *Dunn index.* The Dunn index is calculated by dividing the minimum distance between points in different clusters over the maximum distance between points in the same cluster. This value ranges from zero to infinity, where larger values indicate better quality clusters.

- *Silhouette width.* The silhouette width is calculated by averaging all the silhouette values for each pathway, where the silhouette value is calculated using the following function:

$$silhouetteVal = \frac{b_i - a_i}{\max(a_i, b_i)},$$

where  $a_i$  is the average distance between the pathway and all other pathways in its cluster, and  $b_i$  is the average distance between the pathway and all pathways in the nearest neighboring cluster. This value ranges from  $-1$  to  $1$ , where a value closer to one indicates better quality clusters.

- *Connectivity.* Connectivity is calculated by finding the  $n$  closest pathways to each pathway, then adding  $\frac{1}{n}$  to the total connectivity value if the  $n$ th closest pathway is in a different cluster than the first pathway. This value ranges from zero to

infinity, where smaller values indicate better quality clusters. For all clustering results, connectivity was calculated using  $n = 10$  neighbors. The connectivity was then divided by the total number of pathways to get the adjusted connectivity score.

- *Stability.* Stability is calculated by omitting all the pathways in one cluster, then reclustering the remaining pathways into  $N-1$  clusters, where  $N$  is the original number of clusters. The percentage of pathways that are clustered together in the same way as they were before omitting one cluster can then be calculated. The resulting percentages of pathways that remain in the same clustered groups are averaged across the different omissions to get the stability score.

**External evaluation.** Three external clustering measures were used to assess how well clusters matched up with known literature groupings of the biosynthesis pathways: cluster purity, Normalized Mutual Information (NMI), and adjusted Rand index [18]. The adjusted Rand index is the most strict measure, evaluating on a pairwise-basis if pathways are incorrectly grouped. NMI penalizes more on the cluster-level, while cluster purity falls in between the two other measures.

All external measures depended on having known labels for pathways. Lysine, isoleucine, and 3-HP pathways were assigned five, four, and seven different labels respectively as described in the previous “Metabolic Pathways” subsection of this paper (Figures 1–3).

- *Cluster purity.* Cluster purity was calculated as described in [18]:

$$purity(P, L) = \frac{1}{N} \sum_k \max_j |p_k \cap l_j|,$$

where  $N$  is the number of total pathways,  $P = [p_1, p_2, \dots, p_k]$  is the set of predicted pathway clusters,  $L = [l_1, l_2, \dots, l_j]$  is the set of literature-based pathway clusters, and  $p_k$  and  $l_j$  represent the set of pathways included in predicted cluster  $k$  and in literature-based cluster  $j$  respectively.

There are two features of cluster purity that could bias evaluations: (1) it is not sensitive to small clusters, and (2) it is biased towards having a larger number of clusters. Since the number of clusters is held constant in this study, we only consider the first feature.

- *Purity based on percentage.* The cluster purity measure based on percentage was devised to correct for the original cluster purity’s insensitivity to small clusters:

$$percPurity(P, L) = \frac{1}{k} \sum_k \max_j \frac{|p_k \cap l_j|}{|l_j|},$$

where the symbols are the same as described for the original purity value calculation. Having a purity value based on percentage allows us to weight each cluster equally, even though the cluster groups may vary in size.

- *Normalized mutual information (NMI).* NMI is a measurement of the certainty that the pathway belongs to a specific literature-based cluster if it was selected from any given predicted cluster. The value is divided by an entropy value, which gets larger with increasing number of clusters.

• *Adjusted Rand index.* The adjusted Rand index is a measure of simple accuracy. To use the adjusted Rand index, the literature-based clusters were treated as the true clustering result.

## 4 Implementation

**Distances.** The Levenshtein distance algorithm [27] was implemented in Python. Jaccard distance was calculated using the R “dist” function. The  $n$ -gram graph distance was calculated using the JInsect program, a Java implementation of Giannakopoulos et al. [9].

**Clustering methods.** Clustering calculations and analyses were performed using the statistical program R [23]. For agglomerative hierarchical clustering, pathway distance matrices were passed to the “agnes” function in the “cluster” package [17]. To obtain a specified number of clusters, the “cutree” function was used on the result. The “kmeans” function in the “stats” package and the “pam” function in the “cluster” package were used for  $k$ -means clustering and  $k$ -medoids clustering respectively.

**Evaluation measures.** The Dunn index, silhouette width, and connectivity of clusters were analyzed using the “dunn”, “silhouette”, and “connectivity” functions in the “cIValid” package [2]. NMI was calculated in R using the “clue” package [12]. Adjusted Rand index was calculated in R using the “mclust” package [8].

## 5 Results

**Internal measures.** Based on the internal clustering measures, clusters using  $n$ -grams distance appeared to be higher quality than the other distance measures. However, there was no one clustering method or distance measure that consistently yielded the highest quality clusters. The 3-HP pathways appeared to have the lowest quality clusters among the three biosynthesis pathways.

No adjusted connectivity values exceeded 0.8 across all pathways, clustering methods, and distance measures. Clusters using  $n$ -gram distance seemed to have the lowest adjusted connectivity values across all pathways. For all but one instance, clusters obtained using  $k$ -medoids had the highest adjusted connectivity value within any given distance measure (Table 1).

Dunn index for all clustering methods and distances were low (average value = 0.27). Dunn index only exceeded 0.5 for isoleucine clusters using hierarchical clustering and  $n$ -grams distance. The average Dunn index for 3-HP clusters (0.2) was lower than those of lysine (0.3) and isoleucine (0.31). Clusters using  $n$ -grams distance appeared to have the highest Dunn index values across pathways. Silhouette width across pathways was also low (average value = 0.3), with the exception of  $k$ -medoid and  $k$ -means clustering combined with

Table 1: Internal evaluation measures for lysine (A), isoleucine (B), and 3-HP (C) biosynthesis pathways. Values highlighted in green indicate that the clusters generated by the corresponding clustering method and distance measure resulted in the best value for that given internal measure, whereas values highlighted in red indicate the clusters had the worst value.

A. Lysine		Adjusted			
Clustering method	Distance	connectivity	Dunn	Silhouette	Stability
hierarchical	Levenshtein	0.306	0.20	0.33	0.92
k-medoids	Levenshtein	0.530	0.26	0.84	0.96
k-means	Levenshtein	0.001	0.29	0.86	0.96
hierarchical	Jaccard	0.204	0.40	0.28	0.75
k-medoids	Jaccard	0.511	0.28	0.23	0.79
k-means	Jaccard	0.301	0.28	0.24	0.88
hierarchical	$n$ -gram	0.119	0.32	0.36	0.71
k-medoids	$n$ -gram	0.258	0.32	0.94	0.94
k-means	$n$ -gram	0.001	0.39	0.86	0.96

B. Isoleucine		Adjusted			
Clustering method	Distance	connectivity	Dunn	Silhouette	Stability
hierarchical	Levenshtein	0.493	0.25	0.12	0.70
k-medoids	Levenshtein	0.615	0.25	0.25	0.25
k-means	Levenshtein	0.739	0.25	0.19	0.80
hierarchical	Jaccard	0.511	0.32	0.21	0.81
k-medoids	Jaccard	0.658	0.31	0.15	0.77
k-means	Jaccard	0.464	0.32	0.16	0.94
hierarchical	$n$ -gram	0.096	0.53	0.29	0.89
k-medoids	$n$ -gram	0.067	0.20	0.39	0.93
k-means	$n$ -gram	0.094	0.32	0.23	0.89

C. 3HP		Adjusted			
Clustering method	Distance	connectivity	Dunn	Silhouette	Stability
hierarchical	Levenshtein	0.240	0.22	0.13	0.37
k-medoids	Levenshtein	0.787	0.20	0.11	0.93
k-means	Levenshtein	0.693	0.22	0.12	0.69
hierarchical	Jaccard	0.311	0.12	0.10	0.49
k-medoids	Jaccard	0.683	0.27	0.09	0.89
k-means	Jaccard	0.358	0.11	0.09	0.83
hierarchical	$n$ -gram	0.348	0.21	0.16	0.19
k-medoids	$n$ -gram	0.366	0.21	0.16	0.96
k-means	$n$ -gram	0.354	0.21	0.16	0.76

Levenshtein and  $n$ -gram distances for lysine biosynthesis pathways. The average Silhouette width was also lower for 3-HP (0.12) than for those for lysine (0.55) and isoleucine (0.22). For Silhouette width and to a lesser extent Dunn index, clustering using Jaccard distance had the lowest values across the three pathways.

Lysine clusters using Levenshtein distance showed the most stability, with values above 0.9 across all clustering methods. Isoleucine clusters using  $n$ -gram distance showed the most stability, with values close to 0.9. For 3-HP clusters, there was no one distance measure that resulted in high stability values; however, clusters obtained using hierarchical clustering had significantly lower stability values (Table 1).

Across all pathways, hierarchical clustering yielded clusters with the lowest stability values on average compared to  $k$ -medoids and  $k$ -means clustering.

Table 2: External evaluation measures for lysine (A), isoleucine (B), and 3-HP (C) biosynthesis pathways. Values highlighted in green indicate that the clusters generated by the corresponding clustering method and distance measure resulted in the best value for that given external measure, whereas values highlighted in red indicate the clusters had the worst value.

#### A. Lysine

Clustering method	Distance	Standard purity	Purity, percentage	Adjusted Rand index	NMI
hierarchical	Levenshtein	0.953	0.790	0.931	0.899
k-medoids	Levenshtein	0.720	0.664	0.561	0.657
k-means	Levenshtein	0.839	0.790	0.789	0.851
hierarchical	Jaccard	0.928	0.665	0.873	0.834
k-medoids	Jaccard	0.699	0.597	0.556	0.633
k-means	Jaccard	0.653	0.465	0.321	0.407
hierarchical	n-gram	0.627	0.600	0.441	0.571
k-medoids	n-gram	0.631	0.606	0.553	0.670
k-means	n-gram	0.610	0.501	0.294	0.390

#### B. Isoleucine

Clustering method	Distance	Standard purity	Purity, percentage	Adjusted Rand index	NMI
hierarchical	Levenshtein	0.844	0.667	0.526	0.618
k-medoids	Levenshtein	0.403	0.441	-0.034	0.093
k-means	Levenshtein	0.435	0.470	0.066	0.246
hierarchical	Jaccard	0.532	0.500	0.172	0.451
k-medoids	Jaccard	0.409	0.413	-0.042	0.098
k-means	Jaccard	0.484	0.500	0.165	0.449
hierarchical	n-gram	0.575	0.675	0.229	0.478
k-medoids	n-gram	0.489	0.674	0.059	0.218
k-means	n-gram	0.489	0.578	0.080	0.280

#### C. 3-HP

Clustering method	Distance	Standard purity	Purity, percentage	Adjusted Rand index	NMI
hierarchical	Levenshtein	0.692	0.545	0.458	0.562
k-medoids	Levenshtein	0.438	0.482	0.242	0.398
k-means	Levenshtein	0.505	0.528	0.282	0.476
hierarchical	Jaccard	0.538	0.615	0.293	0.567
k-medoids	Jaccard	0.411	0.538	0.254	0.461
k-means	Jaccard	0.488	0.545	0.259	0.478
hierarchical	n-gram	0.490	0.386	0.192	0.368
k-medoids	n-gram	0.497	0.487	0.275	0.451
k-means	n-gram	0.519	0.450	0.290	0.449

**External measures.** Across all clustering methods and distances, lysine pathway clusters scored the highest across most external measures (Table 2). The combination of hierarchical clustering and Levenshtein distance had the highest values for most external measures across pathways. For lysine pathways, clusters obtained using Levenshtein distance matched with expected pathway groupings better than both Jaccard and  $n$ -gram distances. For isoleucine pathways, Levenshtein distance performed well only with hierarchical clustering; across all three distance measures, hierarchical clustering yielded the best scores for isoleucine pathways. For 3-HP, no one distance or clustering measure appeared to perform well across all external evaluation measures. Hierarchical clustering with Levenshtein and Jaccard distances had the slightly higher external evaluation scores, while hierarchical

Table 3: Spearman correlations between internal and external evaluation measures. Correlations with asterisks are statistically significant ( $p < 0.05$ ).

External \ Internal	Connectivity	Dunn index	Silhouette width	Stability
Purity	-0.474*	0.202	0.450*	0.066
Purity by percentage	-0.508*	0.116	0.394*	0.250
adjustedRandIndex	-0.306	0.070	0.311	0.119
NMI	-0.233	0.103	0.202	0.092

clustering with  $n$ -gram distance had lower scores.

#### Correlation between external and internal measures.

Spearman’s correlation coefficients [19] were calculated between all external and internal evaluation methods across all pathways, distance measures, and clustering methods. There was a statistically significant negative correlation between both standard and percentage purity values and connectivity, and a significant positive correlation between the two purity values and silhouette width. This means that good clusters that matched external labeling based purity scores tended to also be evaluated as good clusters based on internal connectivity and silhouette width scores. Similarly in the case of bad clusters, the purity values and the connectivity and silhouette width scores tended to agree. All other correlations had  $p$  values greater than 0.05 (Table 3).

## 6 Discussion

Hierarchical clustering with Levenshtein distance performed the best overall in terms of external evaluation measures. However, this combination did not perform as well for the internal evaluation measures. One possible reason for the low internal evaluations scores is the differing separation levels of the clusters. Of the known lysine biosynthesis pathways, four of the expected labeled groups go through meso-2,6-diaminopimelate (DAP, see Figure 1). Pathways in these clusters have smaller pairwise distances between each other, especially when using Levenshtein distance since only a few reactions need to be changed to convert from one labeled group to another. As a result, these clusters are not as well separated, which results in a lower internal evaluation score. The clusters using  $n$ -grams distance may have higher internal evaluation scores because the  $n$ -grams captures more of the differences between these closely related clusters, so the pairwise differences and thus the separation between these clusters is larger. The higher internal evaluation scores suggest that  $n$ -grams distances are more sensitive to smaller variations between pathways and may be better at separating pathways that share a common set of reactions than Levenshtein distance.

For all three biosynthesis pathways, the Dunn index never exceeded one, which suggests that on average the minimum distance between two pathways in different clusters was

smaller than the maximum distance between two pathways in the same cluster. However, the Silhouette width for all pathways was never negative, indicating the average distance between pathways within a cluster was always smaller than the average distance between pathways in different clusters. For lysine, the Dunn index between the AAA and all of the DAP clusters combined was no less than 0.8 across most distance measures and clustering techniques. However, since the Dunn index between each of the four DAP clusters were much smaller, the total Dunn index between all lysine clusters did not exceed 0.55 and mostly fell under 0.3 across all clustering and distance measures. Having a distance measure that can better separate pathways with shared reactions would improve the internal evaluation measures for the four DAP pathway groups in lysine. In the cases where Dunn index and Silhouette width did not appear tightly correlated (i.e., for lysine clusters using  $k$ -mean and  $k$ -medoids clustering and Levenshtein and  $n$ -grams distance), the difference correctly indicates that there are at least a few “outlier” pathways in the clusters with large pairwise distances between them, causing the Dunn index to be unexpectedly lower or higher relative to the Silhouette width.

For simple clustering cases like the lysine biosynthesis pathway where there were two distinct pathway groups defined in the literature, using hierarchical clustering and Levenshtein distance yielded well defined clusters that agreed with the expected labeling. However, for more complex cases like 3-HP which had several overlapping pathways (Figure 3), clusters were less defined for all clustering methods and distances. Though simple clustering methods are effective in grouping pathways with well-defined differences, the results for 3-HP and isoleucine suggest that simple clustering is not an effective solution for pathways with many shared and interweaving reaction chains. A more hierarchical approach based on the underlying topology of the metabolic network may yield better pathway grouping results.

Significant correlations were found between external purity scores and internal connectivity and Silhouette width scores. This suggests that connectivity and Silhouette width may provide a better assessment of the quality of pathway clusters in the absence of external validation than other internal evaluation measures. This finding is in agreement with a conclusion of Wiwie et al. [28], who tentatively recommended silhouette values as the best internal measure of clustering quality when tested against 24 biomedical data sets, but noted it is not a replacement for external indices when available. It would be interesting to test if connectivity and Silhouette width consistently show a significant correlation with external measures across other biosynthesis pathways besides the three pathways evaluated in this study.

When thousands of possible pathway solutions are found (i.e., 3-HP), these pathways share a few common intermediate compounds. These intermediate compounds are often bottleneck points, where several different series of reactions exist that produce this compound and result in the variations

between pathway solutions. Focusing on variations found between series of bottleneck points can improve distance measures between pathways. Iteratively clustering pathways based on variation between bottleneck points between the starting and ending compound can help limit pathway comparison to a shorter segment of the reactions, making it easier to group pathways. Accounting for bottleneck points in pathways or making adjustments to the distance measure based on the pathways’ underlying biochemical features could improve clustering in the future.

Hierarchical clustering has been used widely for many clustering problems. Since hierarchical clustering can be used with any distance measure, it is ideal for problems that require the use of non-triangular distances that do not make sense in a Euclidean space. New ideas for improving distance measures for metabolic pathways could potentially be applied more generally to classifying items with similar underlying structure, including grouping documents [30], clustering protein complexes from protein interaction networks and species from metagenomic DNA [21], and grouping protein and gene sequence data. Gaining a better understanding of grouping metabolic pathways can provide additional insight into similar clustering problems. Distance measures could even be learned to more closely emulate expert knowledge and classification [29]. However, more experiments are needed to confirm that these observations about pathways in this study can be more generally applied to other metabolic pathways. Additionally, our results were evaluated based on pathway grouping found in current literature and expert knowledge, which itself is biased by existing tools. In the future, it would be better to have a way to identify if a cluster with high intrinsic quality also exhibits a significant grouping beyond the existing expert knowledge.

In conclusion, though there was no one combination of clustering methods and distances that performed well across all three example biosynthesis pathways, some clustering methods and distance measures (i.e., hierarchical clustering and Levenshtein distance) resulted in clusters that matched the expected external labeling found in the literature. This suggests that an automated method for grouping pathways is possible given a more directed approach, and a classification method that factors in the pathways’ unique topology (i.e., clustering using more fine-tuned distance measures) may be a promising direction.

## 7 Acknowledgements

This work has been supported in part by NSF DBI-1262491 and the NSF Graduate Fellowship awarded to Sarah Kim. The authors would like to sincerely thank Panagiotis Eustratiadis for his time and generous support in discussing the application of the  $n$ -gram distance algorithm and for developing the tools required to specifically calculate pairwise pathway distances based on the  $n$ -gram graph distance.

## References

- [1] Masanori Arita. Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, 8(1-2):109–125, April 2000.
- [2] Guy Brock, Vasyi Pihur, Susmita Datta, and Somnath Datta. cIValid: An R package for cluster validation. *Journal of Statistical Software*, 25(4):1–22, 2008.
- [3] Ron Caspi, Tomer Altman, Richard Billington, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471, 2014.
- [4] Didier Croes, Fabian Couche, Shoshana J. Wodak, and Jacques van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1):222–36, February 2006.
- [5] David Croft, Antonio Fabregat Mundo, Robin Haw, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue):D472–7, January 2014.
- [6] Martin Ester, Hans peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.
- [7] Karoline Faust, Pierre Dupont, Jérôme Callut, and Jacques van Helden. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, 26(9):1211–8, May 2010.
- [8] Chris Fraley, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca. *mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation*, 2012.
- [9] George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. Summarization system evaluation revisited. *ACM Transactions on Speech and Language Processing*, 5(3):1–39, October 2008.
- [10] Allison P. Heath, George N. Bennett, and Lydia E. Kavrakli. Finding metabolic pathways using atom tracking. *Bioinformatics*, 26(12):1548–55, June 2010.
- [11] Christopher S. Henry, Linda J. Broadbelt, and Vassily Hatzimanikatis. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnology and Bioengineering*, 106(3):462–73, June 2010.
- [12] Kurt Hornik. *clue: Cluster ensembles*, 2015. R package version 0.3-50.
- [13] Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.
- [14] Xinglin Jiang, Xin Meng, and Mo Xian. Biosynthetic pathways for 3-hydroxypropionic acid production. *Applied Microbiology and Biotechnology*, 82(6):995–1003, April 2009.
- [15] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.
- [16] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(Database issue):D199–205, January 2014.
- [17] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version 2.0.3.
- [18] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.
- [19] John H McDonald. *Handbook of Biological Statistics*, volume 3. Sparky House Publishing Baltimore, MD, 2014.
- [20] Geoffrey J. McLachlan and Kaye E. Basford. Mixture models: Inference and applications to clustering. *Applied Statistics*, 1988.
- [21] Saket Navlakha, James White, Niranjan Nagarajan, Mihai Pop, and Carl Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. *Journal of Computational Biology*, 17(3):503–16, March 2010.
- [22] F. J. Planes and J. E. Beasley. An optimization model for metabolic pathways. *Bioinformatics*, 25(20):2723–9, October 2009.
- [23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [24] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and Dietmar Schomburg. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, 21(7):1189–93, April 2005.
- [25] Alberto Rezola, Luis F de Figueiredo, M Brock, Jon Pey, Adam Podhorski, Christoph Wittmann, Stefan Schuster, Alexander Bockmayr, and Francisco J Planes. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, 27(4):534–40, February 2011.
- [26] Maurice Scheer, Andreas Grote, Antje Chang, et al. BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, 39(Database issue):D670–6, January 2011.
- [27] R. A. Wagner and M. J. Fisher. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173, January 1974.
- [28] Christian Wiwie, Jan Baumbach, and Richard Röttger. Comparing the performance of biomedical clustering methods. *Nature Methods*, 12(11):1033–1038, September 2015.
- [29] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State University*, 2, 2006.
- [30] Ying Zhao, George Karypis, and Usama Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168, March 2005.