Unsupervised Machine Learning or Clustering of Sequences
Drosophila : cytochrome oxidase subunit I (COI) and alcohol dehydrogenase (Adh) clustering patterns

Veedhi Solanki

October 27, 2023

INTRODUCTION:

Unsupervised machine learning, with a focus on clustering, has become a powerful tool for interpreting the evolutionary relationships and genetic diversity among taxonomic groups. As in study by Liu et al., 2018, shows Incorporating a coevolutionary method and hierarchical clustering has shown to yield more robust gene clusters, revealing coherent evolutionary patterns and functional relationships among genes. There are variety of clustering methods available such as Edge-betweenness algorithm, modularity maximization method and many more where choosing one depend on the type of dataset such that in a study by Leger et al., 2015 focused on identifying optimal methods for detecting highly interactive species subgroups within bipartite networks, which offers precise recommendations for selecting suitable clustering techniques based on network type, and providing readily accessible simulation code for future comparative analyses. Researchers can find unique clustering patterns in genetic sequences by using this method, which sheds light on the significance of various genes as markers for classification. It has been difficult previously without machine learning to look at clustering patterns efficiently such that time consuming methods like 2D image clustering were used as in Peng et al., 2006. This study involves analysis of possible differences in clustering patterns with a focus on the taxonomic group Drosophila, using the genes cytochrome oxidase subunit I (COI) and alcohol dehydrogenase (Adh). Knowing how these genes cluster within the same taxonomic group can help reveal the functional limitations and evolutionary factors affecting genomic sequences. Therefore, the primary objective of this study is to investigate whether two different genes COI and Adh, exhibit similar or different clustering patterns within the Drosophila genus. It will be also interesting to compare clustering patterns of these two genes because COI is a mitochondrial gene while Adh is found on chromosome 4. So, it will be important finding whether somatic genes are good biomarkers to identify revolution within a taxon specifically for genus Drosophila. In addition, this study also predicts any variations in evolutionary pressures—such as functional limitations—that could affect how these genes cluster.

Given that COI and Adh are vital genes involved in important physiological processes. Adh is used in alcohol metabolism while COI is involved in electron transport chain in mitochondria. Additionally, COI is a very essential biomarker used to investigate phylogenies. So, the genes being used here are very contrasting in their utilization in looking at evolutionary patterns. It is speculated that they would exhibit different clustering patterns in the Drosophila genus. Due to the contrasting nature and uses of these two genes in understanding evolutionary patterns, the primary hypothesis is that there will be substantial variation in the clustering patterns of COI and Adh gene sequences throughout the Drosophila genus, indicating a range of evolutionary pressures and functional roles for various genes. This variance would also suggest the presence of particular gene clusters that would be better suited as markers to categorize different subgroups within the genus. However, it is also possible that there are notable similarities between gene sequences of somatic gene Adh and mitochondrial gene COI and so

similar clustering patterns. This can be an indication to a conserved evolutionary pressure and functional constraints across genes and implying that they are suitable as reliable markers for classification applications. These contrasting hypotheses present an interesting opportunity to investigate the ways in which distinct genes within the Drosophila genus cluster together. The degree of similarity or dissimilarity between the genes can be examined using this approach, providing insights into their evolutionary relationships and possible implications for classification applications. The findings of this study will have a big impact on how we understand genetic diversity and evolutionary processes in the Drosophila taxonomic group.

RESULTS AND DISCUSSION:

The sequence data for the genes COI and Adh for the genus were obtained from the NCBI as it is one of the best sources for it. Gene length restriction was also applied in the search to restrict the sequences to the real gene and avoid getting false positive sequences. Moreover, the data was obtained in small batches with for loop which were then merged and also converted to dataframe to better visualize and work with throughout the project. The code for search term used is as below:

```
search_term_COI <- "(Drosophila[Organism] AND COI[Gene] AND 600:700[SLEN])"
search_term_Adh <- "(Drosophila[Organism] AND Adh[Gene]AND 600:700[SLEN])"
search_COI_2 <- entrez_search(db = "nuccore", term = search_term_COI, retmax = maxHits_COI, use_history = T)
search_Adh_2 <- entrez_search("nuccore", term = search_term_Adh, retmax = maxHits_Adh, use_history = T)
```

Furthermore, data was filtered such that first removing the rows with not available sequence data then then N or dashes at the beginning or end of sequence data were removed. Then ensuring that the count is less than or equal to a calculated threshold based on the variable missing.data of 0.01. Finally the rows were filtered based on the count of characters in the sequences within a range centered around the median of the character count, with the range defined by the variable length.var which was set to 50. Additionally, the comprehensive title column was divided into separate columns such as accession, voucher, gene description, sequence description and species name. These all allowed to get quality data for further analysis. The sample code for COI followed is as shown below:

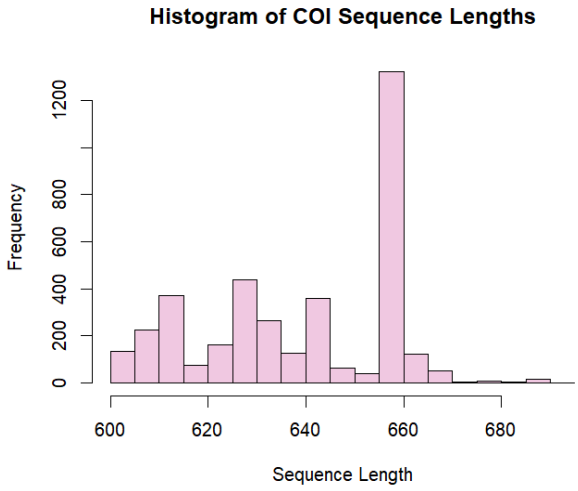
```
#For COI:
dfCOI_1 <- dfCOI %>%
  filter(!is.na(COI_Sequence)) %>%
  mutate(COI_Sequence = str_remove_all(COI_Sequence, "^N+|N+$|-")) %>%
  filter(str_count(COI_Sequence, "N") <= (missing.data * str_count(COI_Sequence))) %>%
```

```
    filter(str_count(COI_Sequence) >= median(str_count(COI_Sequence)) - length.var &
str_count(COI_Sequence) <= median(str_count(COI_Sequence)) + length.var)
dfCOI_1 <- separate(dfCOI_1, COI_Title, into = c("Accession", "Voucher", "Gene_Description",
"Sequence_Description"), sep = " ", extra = "merge")
```

Before doing the further analysis it was very essential to get an overview of sequence length distribution and to make sure data being used is valid as well as if the applied filters worked. So, the histogram was produced from the following sample code for COI:

```
my_colors <- brewer.pal(8, "Pastel2")
# Histogram for COI Sequence
hist(nchar(dfCOI_1$COI_Sequence), col = my_colors[4],
    main = "Histogram of COI Sequence Lengths",
    xlab = "Sequence Length",
    ylab = "Frequency")
```

A)



B)

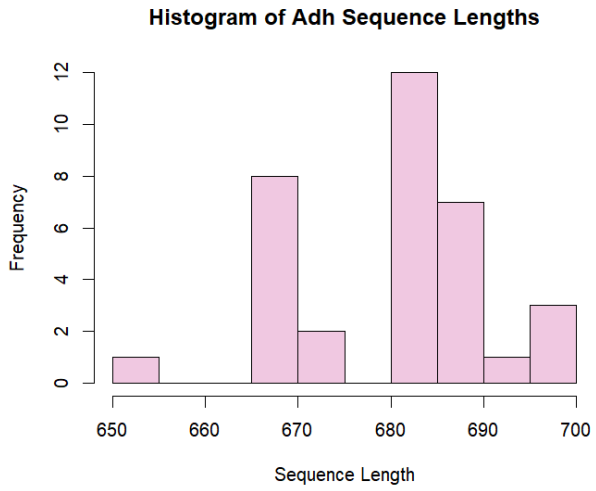


Figure 1: It demonstrates the sequence length distribution and frequency for genes **A)** COI and **B)** Adh that are being used for the further analysis.

The lengths of distribution were within the range as from **Figure 1**. The COI sequences are broadly distributed but highest frequency at 680 base pairs. Similarly, Adh sequences also have the highest frequency at 650 and 680-690 base pairs.

Sequences were aligned using the code below with appropriate penalty for gaps, however, the next step which involved performing pairwise distance alignment matrix, needed to remove gaps from the alignment. So, after getting an overview of the alignment with gaps, another alignment was performed where gaps were removed and code of which for COI is as follow but same was used for Adh as well:

```
sequence_data_COI <- readDNAStringSet("COI_alignment.fasta")
sequence_data_COI_XStringSet <- as(sequence_data_COI, "XStringSet")
# Remove gap characters from the sequences
sequence_data_COI_no_gaps <- gsub("-", "", sequence_data_COI_XStringSet)
# Convert back to DNAStringSet
sequence_data_COI_no_gaps <- DNAStringSet(sequence_data_COI_no_gaps)
# Perform sequence alignment on the data without gaps
alignment_COI <- AlignSeqs(sequence_data_COI_no_gaps)
BrowseSeqs(alignment_COI)
```

The sequences for both genes were well aligned and so the pairwise distance matrix was calculated for each gene and converted to the dataframe for better visualization of data. Here is example code for the pairwise distance matrix of COI and similarly conducted for Adh:

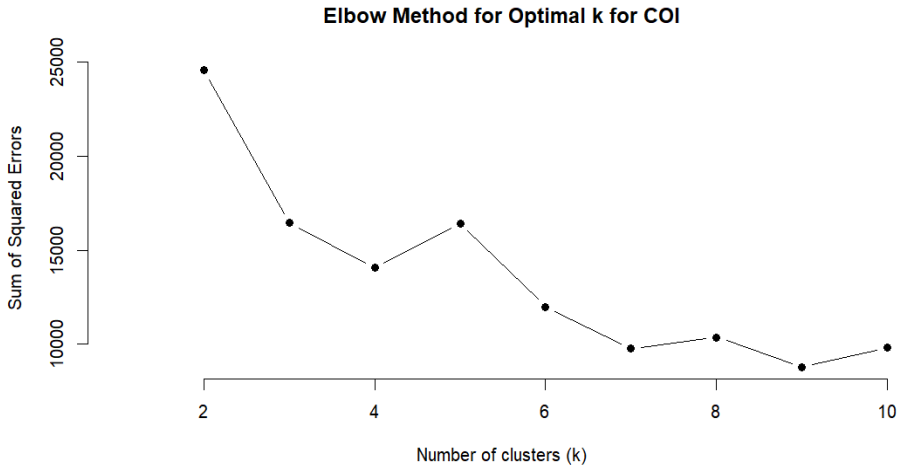
```
# Calculate pairwise distances
dist_mat_COI <- dist.dna(sequence_matrix_COI)
# Convert the distance matrix to a data frame for better visualization
dist_df_COI <- as.data.frame(as.matrix(dist_mat_COI))
# Convert the distance matrix to a data frame
dist_df_COI <- as.data.frame(as.matrix(dist_mat_COI))
rownames(dist_df_COI) <- colnames(dist_df_COI)
```

Pairwise distance matrix is a great measure which shows disparities or similarities between sequences of each gene. This helps to make clusters further the study, look at evolutionary relationships and genetic divergence analysis. The comparison of the pairwise distance matrices for COI and Adh can enable to discern any distinct clustering patterns and determine whether these genes exhibit differential clustering behaviors within the Drosophila taxonomic group. Since the matrix is huge, it is essential to visualize it in some way and some options include dendrogram and t-SNE plot. However, the data for the COI gene is huge and so the dendrogram is very hard to visualize, so, t-SNE plot was used. However, the dendrogram for Adh is perfect and results can be compared to the elbow diagram number of clusters.

To decide the number of clusters that are appropriate to choose for the further analysis, an elbow plot was produced. The elbow plot allowed to choose optimal values for the cluster for each gene. Here is the example code for COI gene:

```
# Compute Sum of Squared Errors (SSE) for different values of k : COI
set.seed(123)
sse_COI <- c()
for (i in 2:10) {
  kmeans_out_COI <- kmeans(distanceMatrix_COI, centers = i)
  sse_COI[i] <- kmeans_out_COI$tot.withinss}
# Plot the elbow plot : COI
plot(1:10, sse_COI, type = "b", pch = 19, frame = FALSE,
    xlab = "Number of clusters (k)", ylab = "Sum of Squared Errors",
    main = "Elbow Method for Optimal k")
```

A)



B)

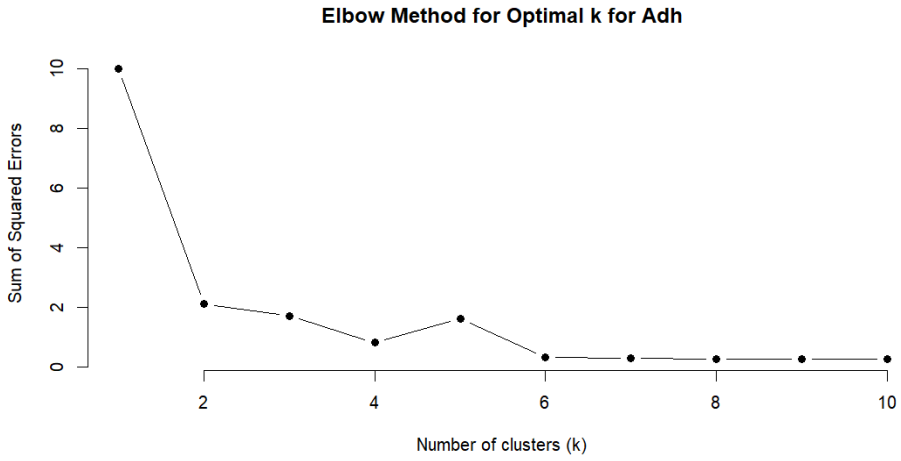


Figure 2: It shows the elbow diagram for **A)** COI and **B)** Adh to choose optimal number of clusters or the value of K for t-SNE plot as we all for Silhouette index.

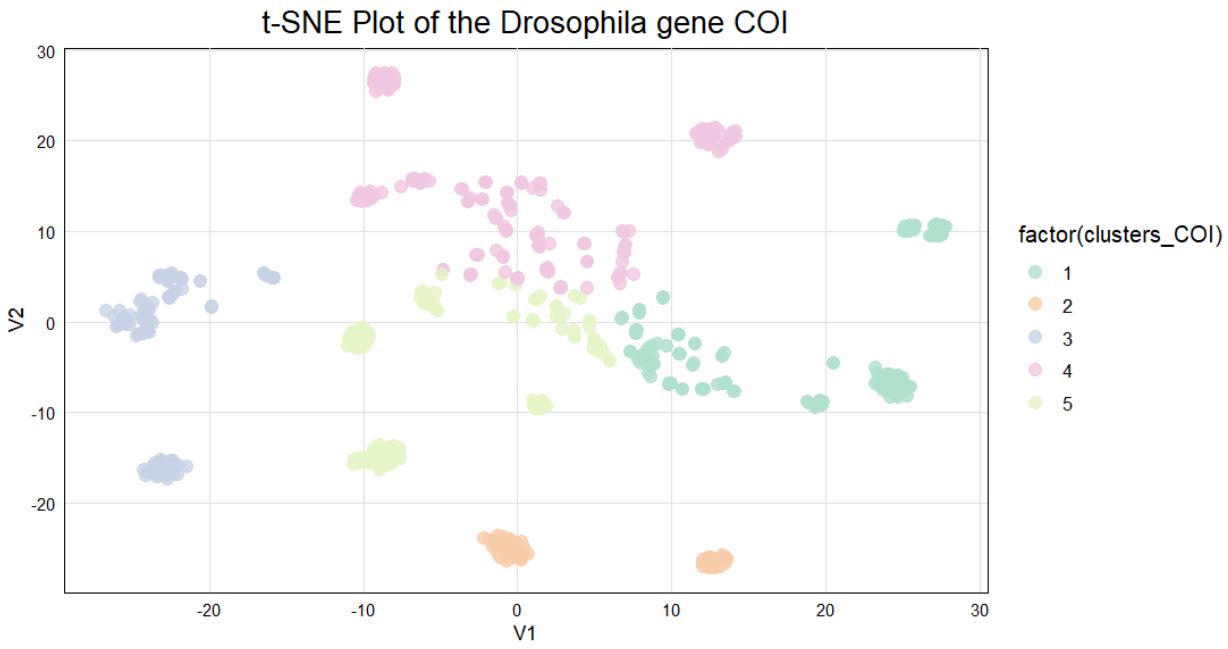
As seen in Figure **2**, the shape is not perfect but it does give an idea of the number of optimal clusters to use for each gene. So, for COI, the elbow point is around 5 while for Adh, it is around 3. So, for further analysis 5 will be the k value for COI and 3 for Adh. The result of the number of clusters from the dendrogram of Adh are comparable to the elbow diagram, as the dendrogram was divided into three main branches at the beginning.

After which, a t-SNE plot was used to look at clustering patterns. The code used for it is as follow for COI:

```
# Run t-SNE on the unique matrix
```

```
tsne_result_COI <- Rtsne(unique_dist_mat_COI)
clusters_COI <- kmeans(tsne_result_COI$Y, centers = 5)$cluster
# Plot the t-SNE result
plot_COI <- ggplot(data = as.data.frame(tsne_result_COI$Y), aes(x = V1, y = V2, color = factor(clusters_COI))) +
  geom_point(size = 3, alpha = 0.8) +
  labs(title = "t-SNE Plot of the Drosophila gene COI") +
  scale_color_brewer(palette = "Pastel2") +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(color = "gray90"),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        axis.text = element_text(color = "black"),
        legend.title = element_text(size = 12, color = "black"),
        legend.text = element_text(size = 10, color = "black"),
        plot.title = element_text(hjust = 0.5, size = 16, color = "black"),
        text = element_text(color = "black"))
plot_COI
```

A)



B)

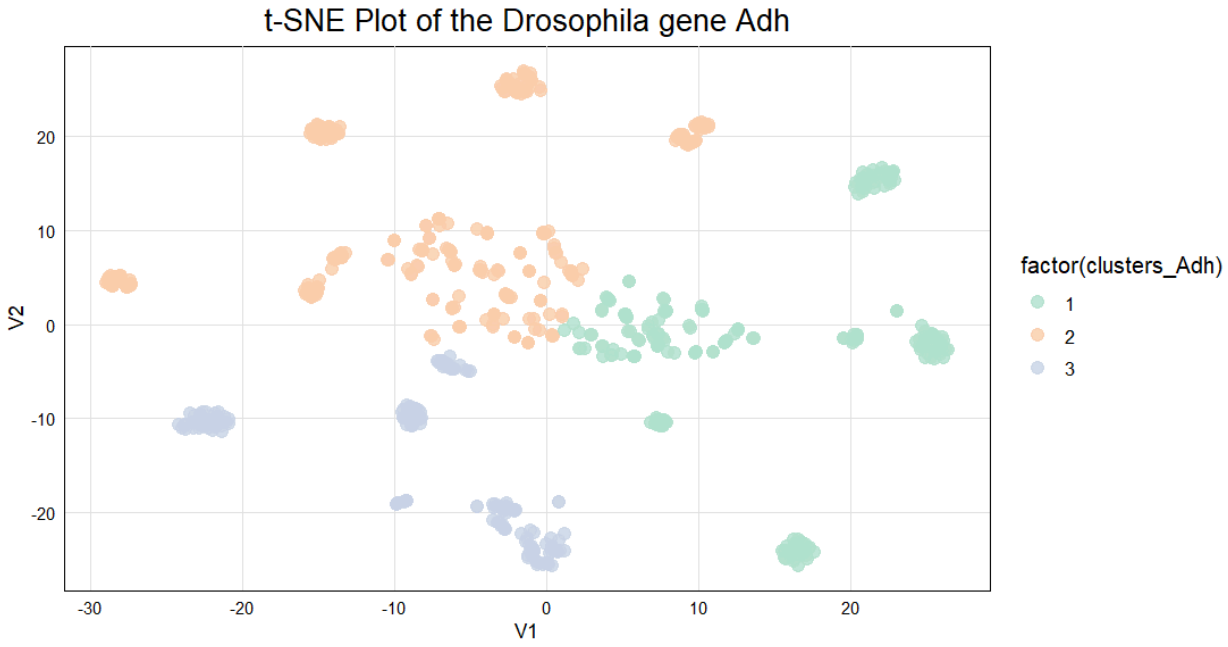


Figure 3: It shows a t-SNEplot for **A)** COI (Clusters = 5) and **B)** Adh (Clusters = 3).

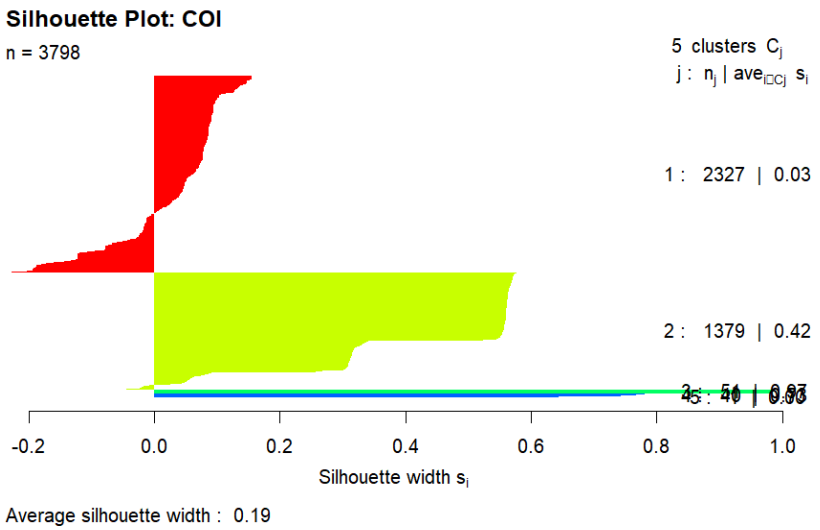
The t-SNE plot used to visualize the pairwise distance matrix showed how similar or distinct sequences are from each other as well as clustering pattern also signified conserved regions in sequences across the genus Drosophila. The axes of the t-SNE plot signifies the comparison of sequences and distance between them. It captures data in high dimension in reduced dimension on plot. As in **Figure 3**, COI gene has five significant clusters while Adh has three significant clusters. Despite the difference in the number of clusters, comparing both figures **A** and **B** more closely reveals that there are some sub-clusters that are similar in both genes. But majorly, looking at it overall, the clustering patterns do differ between the gene COI and Adh.

Finally, the silhouette index was calculated and visualized to compare cluster strength to each other for each gene. Here is the sample code for the COI:

```
COIdistanceMatrix <- as.matrix(distanceMatrix_COI)
COIhclust_object <- hclust(as.dist(COIdistanceMatrix), method = "complete")
k <- 5
COIclusters <- cutree(COIhclust_object, k = k)
sil_COI <- silhouette(COIclusters, dist = COIdistanceMatrix)
summary_sil_COI <- summary(sil_COI)
par(mar = c(5, 4, 4, 8))
```

```
plot(sil_COI, col = rainbow(k), border = NA, main = "Silhouette Plot", cex.names = 0.6)
```

A)



B)

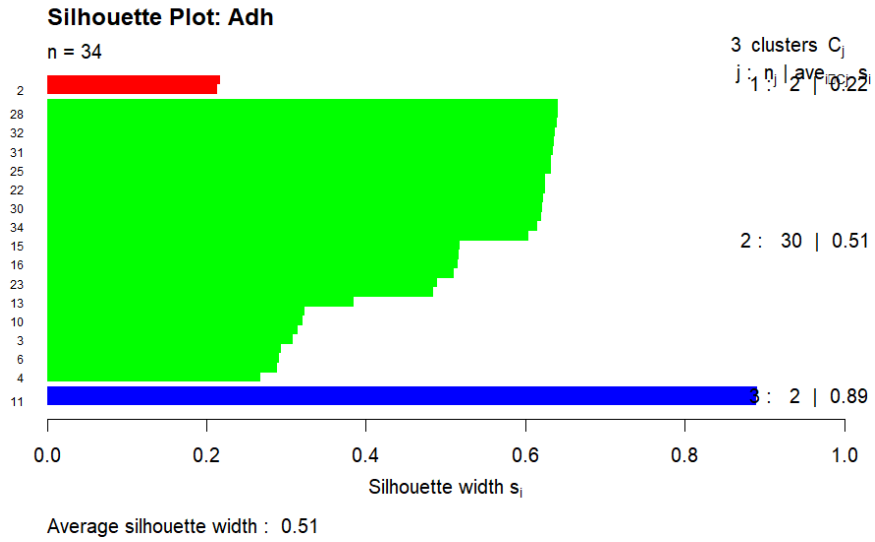


Figure 4: It shows a silhouette plot for **A)** COI (Clusters = 5) and **B)** Adh (Clusters = 3).

The silhouette plot for the genes COI and Adh as shown in **Figure 4**, indicates clustering strengths with individual genes. The cluster one for COI does not have a very high coefficient and it seems to be a misclassified cluster. Additionally, clusters three, four and five are overlapping for COI which is also an indication of misclassification , however, cluster two seemed to be a very great fit. For Adh, plot indicates very high clustering strength in comparison to COI overall, and the cluster two and three are well fit. The Silhouette plot can also be used to

see why some points closer together in the t-SNE plot are of different colors, this is because some clusters here are misclassified and some are very strongly classified.

The analysis of the clustering patterns of the COI and Adh genes in the Drosophila genus revealed distinct clustering behaviors between the two genes. Notable variations in the clustering patterns were shown by the t-SNE plots, suggesting that the somatic Adh gene and the mitochondrial COI gene have different evolutionary paths within the genus Drosophila. The t-SNE and silhouette plots show that although both genes showed clusters indicating conserved regions within the Drosophila genus, there were notable differences in the number of clusters found. These findings imply that within the taxonomic group, the two genes might be exposed to various evolutionary pressures and functional restrictions. Furthermore, the analysis of silhouette plots revealed information about both the benefits and drawbacks of the clusters found within each gene. The COI gene showed certain misclassified clusters, which could indicate that different evolutionary processes are at work on the gene sequences, while the Adh gene showed a stronger overall clustering strength. These results highlight how crucial it is to take into account each gene's distinct properties when utilizing them as molecular markers in classification and evolutionary research.

Overall, the study effectively illustrated how the COI and Adh genes within the Drosophila genus cluster. The project demonstrates the usefulness of unsupervised machine learning methods in clarifying the functional constraints and evolutionary relationships of genes within a taxonomic group, such as silhouette analysis and t-SNE. The contrasting clustering patterns observed between the mitochondrial COI and somatic Adh genes emphasize the need for careful consideration when selecting suitable genetic markers for classification and evolutionary studies. While the findings contribute to a better understanding of the genetic diversity and evolutionary processes in the Drosophila taxonomic group, more research is necessary to fully understand the underlying mechanisms causing these unique clustering behaviors.

Moreover, the results can be influenced by some of the biases and errors encountered in this study. Such that the dataset for COI was a lot larger than Adh which could lead to contrasting result. Moreover, The elbow diagram produced were not perfect so there was a bit of guessing for the exact number of cluster which then influenced downstream analysis. Additionally, genes used to compare were mitochondrial and somatic, it would be very interesting topic to look at two mitochondrial or two somatic genes. The clustering patterns and functional limitations of the COI and Adh genes in the Drosophila genus were the main subjects of the study. But it might have missed other important evolutionary dynamics that could have affected the observed genetic diversity and clustering behaviors within the taxonomic group, like gene duplication events, genetic recombination, and horizontal gene transfer. So, these are some limitations which need improvement further.

REFERENCES:

Hanchuan Peng, Fuhui Long, M. B. Eisen and E. W. Myers, "Clustering gene expression patterns of fly embryos," 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006., Arlington, VA, USA, 2006, pp. 1144-1147, doi: 10.1109/ISBI.2006.1625125.

Leger, J., Daudin, J., & Vacher, C. (2015). Clustering methods differ in their ability to detect patterns in ecological networks. *Methods in Ecology and Evolution*, *6*(4), 474–481. https://doi.org/10.1111/2041-210x.12334

Liu, C., Wright, B., Allen-Vercoe, E., Gu, H., & Beiko, R. (2018, September 1). *Phylogenetic clustering of genes reveals shared evolutionary trajectories and putative gene functions*. Genome biology and evolution. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6130602/

https://lgatto.github.io/IntroMachineLearningWithR/unsupervised-learning.html