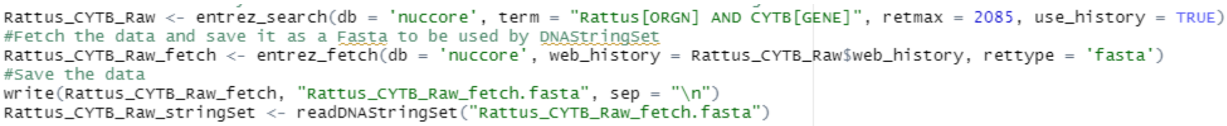
BINF 6210 – Software Tools **Assignment 2** **October 28, 2022**
Student ID: 1111472

Introduction

Phylogeny is an understanding of the evolutionary path of a species. It's a powerful tool that can help us depict biological diversity through intra and inter relationships which can help us glean into the evolution of all organisms on Earth (Baum, 2008). *Rattus* (rats) are organisms which have had close contact with humans throughout history from harbouring deadly diseases to being used for scientific breakthroughs (Yu et al., 2022). Thus, its important to understand the genetic diversity between such organisms as they have had significant impacts on human lives. In this project, I explore the clustering patterns of the COI and CYTB genes of *Rattus* to create a phylogeny to reveal evolutionary divergence. The COI and CYTB genes are mitochondrial genes which have been used in other *Rattus* phylogeny research (Liu et al., 2021) as it has been shown to be a good identification method of genetic diversity within a species. The COI and CYTB genes were extracted from NCBI followed by an alignment and consensus sequence of intersecting *Rattus* species before final clustering based on gene identity. The aim of this study was to determine the best clustering method for the respective genes, and determination of which gene (COI or CYTB) provided a stronger diversification through Silhouette indexing.

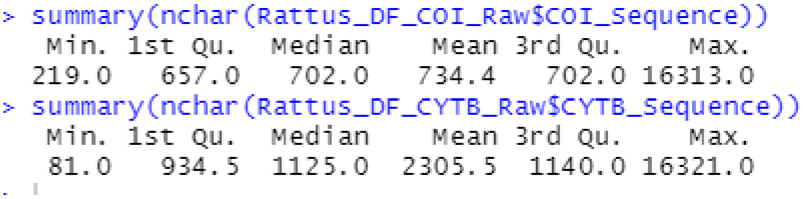
Code Section 1 – Data Acquisition, Exploration, Filtering, and Quality Control

The *Rattus* COI and CYTB gene data was extracted from NCBI using entrez_search() of the nuccore database. It was then formatted into a FASTA file to be used by DNAStringSet for the analysis portion.



```
Rattus_CYTB_Raw <- entrez_search(db = 'nuccore', term = "Rattus[ORGN] AND CYTB[GENE]", retmax = 2085, use_history = TRUE)
#Fetch the data and save it as a Fasta to be used by DNAStringSet
Rattus_CYTB_Raw_fetch <- entrez_fetch(db = 'nuccore', web_history = Rattus_CYTB_Raw$web_history, rettype = 'fasta')
#Save the data
write(Rattus_CYTB_Raw_fetch, "Rattus_CYTB_Raw_fetch.fasta", sep = "\n")
Rattus_CYTB_Raw_stringSet <- readDNAStringSet("Rattus_CYTB_Raw_fetch.fasta")
```

Before processing, an intersect between the *Rattus* species of COI and CYTB was created and only species information that was available in the intersect was used for analysis. Conducting basic summary analysis showed that both the raw datasets had sequence bases that were either too large or too small. For COI, the average bp length has been reported to be around 650 with a variance of +/- 200 bp. For CYTB the average base pair length is roughly 1140bp with high variance (I decided to choose sequences within +/- 400 bp) (Yang et al., 2019). After processing, boxplots were created to show the distribution of COI and CYTB sequence lengths (Figure 1).



```
> summary(nchar(Rattus_DF_COI_Raw$COI_Sequence))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  219.0   657.0   702.0   734.4   702.0 16313.0 
> summary(nchar(Rattus_DF_CYTB_Raw$CYTB_Sequence))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   81.0   934.5  1125.0  2305.5  1140.0 16321.0 
```

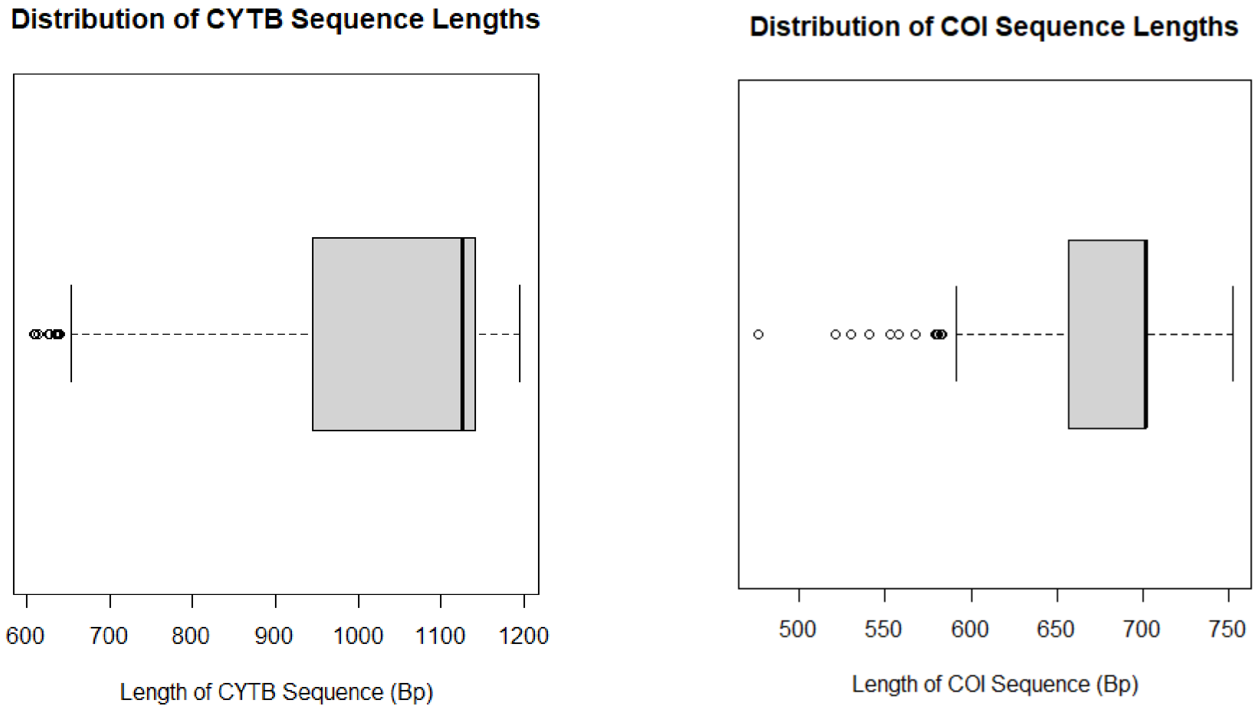
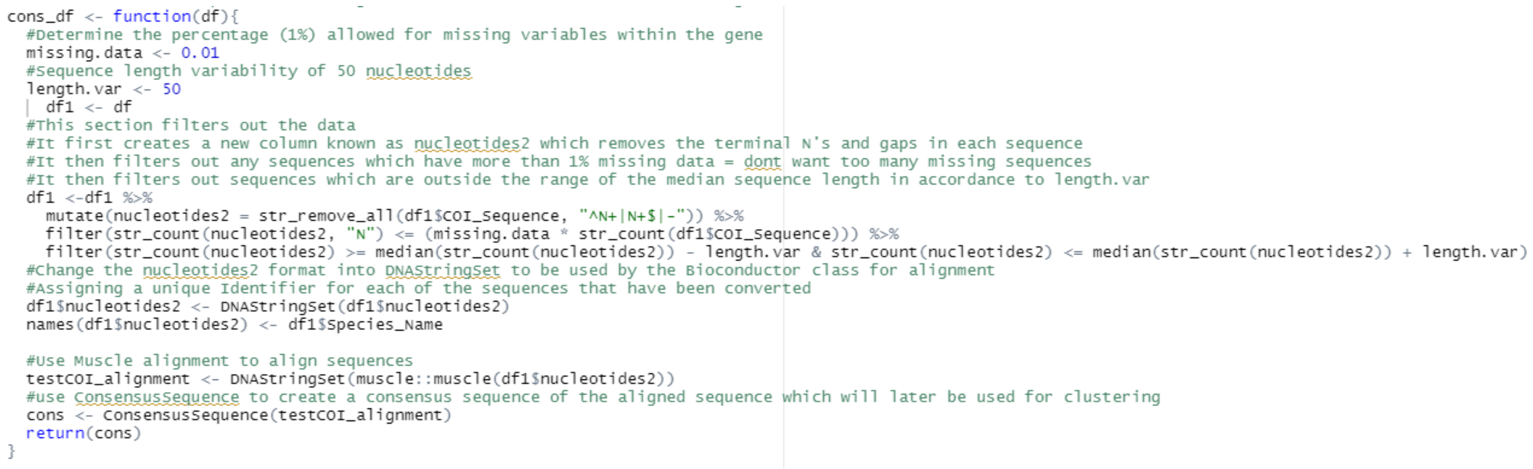


Figure 1. Box plots depicting the length of CYTB and COI sequence after data filtering

Code Section 2 – Main Analysis

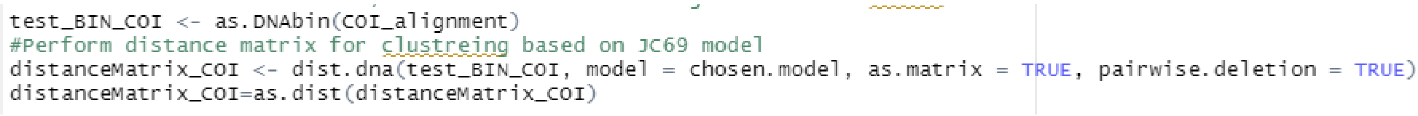
For alignment, both COI and CYTB gene pool were separated by species to create a consensus sequence for each species before alignment and clustering. Sequences were further filtered by trimming of terminal Ns, removal of gaps, removal of sequences with more than 1% missing data (internal Ns) and only chose sequences that fit in a range of 50 nucleotides to the median. The nucleotide sequences were converted into DNAStringSet and then a MUSCLE alignment was done. A consensus sequence was created (using ConsensusSequence) of the aligned COI/CYTB genes for each Rattus subspecies.



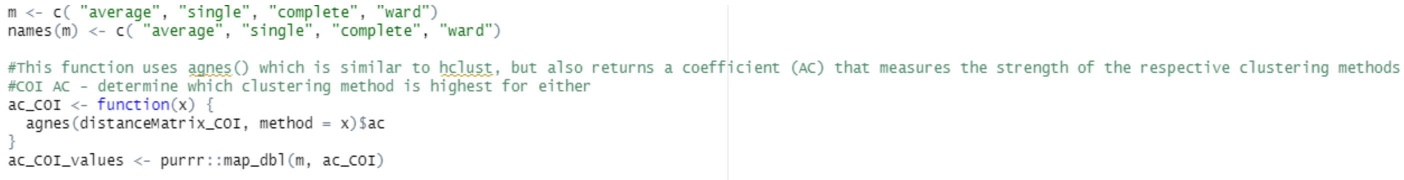
```
cons_df <- function(df){
  #Determine the percentage (1%) allowed for missing variables within the gene
  missing.data <- 0.01
  #Sequence length variability of 50 nucleotides
  length.var <- 50
  | df1 <- df
  #This section filters out the data
  #It first creates a new column known as nucleotides2 which removes the terminal N's and gaps in each sequence
  #It then filters out any sequences which have more than 1% missing data = dont want too many missing sequences
  #It then filters out sequences which are outside the range of the median sequence length in accordance to length.var
  df1 <-df1 %>%
    mutate(nucleotides2 = str_remove_all(df1$COI_Sequence, "^N+|N+$|-")) %>%
    filter(str_count(nucleotides2, "N") <= (missing.data * str_count(df1$COI_Sequence))) %>%
    filter(str_count(nucleotides2) >= median(str_count(nucleotides2)) - length.var & str_count(nucleotides2) <= median(str_count(nucleotides2)) + length.var)
  #Change the nucleotides2 format into DNAStringSet to be used by the Bioconductor class for alignment
  #Assigning a unique Identifier for each of the sequences that have been converted
  df1$nucleotides2 <- DNAStringSet(df1$nucleotides2)
  names(df1$nucleotides2) <- df1$Species_Name

  #Use Muscle alignment to align sequences
  testCOI_alignment <- DNAStringSet(muscle::muscle(df1$nucleotides2))
  #use ConsensusSequence to create a consensus sequence of the aligned sequence which will later be used for clustering
  cons <- ConsensusSequence(testCOI_alignment)
  return(cons)
}
```

After the consensus sequence of each Rattus species was derived for both COI and CYTB genes, a second round of alignment was performed of the consensus sequences.

Clustering was the next step after alignment. Hierarchical clustering approach was used, which required a distance matrix to be made first. The distance model JC69 was chosen as it treated all base frequencies equally, all substitutions being likely. Distance matrixes were made by DNAbin class. Then the optimal clustering method was calculated for COI and CYTB through the agnes() which returns a coefficient (AC) that measures the strength of the clusters. For both COI and CYTB, the Ward clustering method returned the highest coefficient value (0.82).

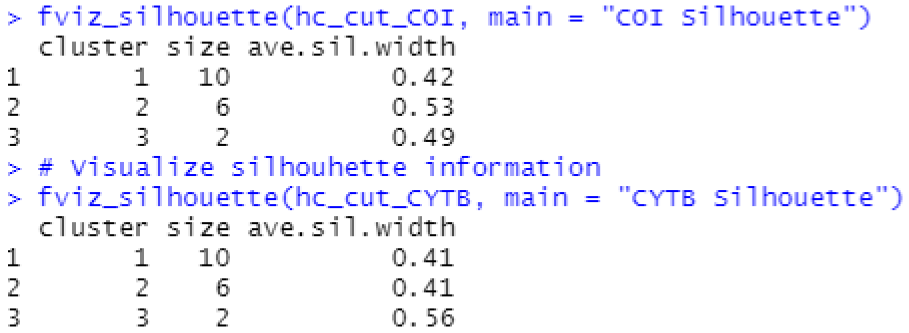


```
test_BIN_COI <- as.DNAbin(COI_alignment)
#Perform distance matrix for clustreing based on JC69 model
distanceMatrix_COI <- dist.dna(test_BIN_COI, model = chosen.model, as.matrix = TRUE, pairwise.deletion = TRUE)
distanceMatrix_COI=as.dist(distanceMatrix_COI)
```



```
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

#This function uses agnes() which is similar to hclust, but also returns a coefficient (AC) that measures the strength of the respective clustering methods
#COI AC - determine which clustering method is highest for either
ac_COI <- function(x) {
  agnes(distanceMatrix_COI, method = x)$ac
}
ac_COI_values <- purrr::map_dbl(m, ac_COI)
```

After optimal cluster method was obtained, a hierarchical cluster was formed for each of the COI and CYTB and a dendogram was plotted to showcase the evolutionary divergence for each respective gene cluster (Figure 2). To determine which gene had the stronger cluster formation, a Silhouette analysis was performed (3 cluster formation). The COI gene had a slightly higher silhouette index (0.46) vs CYBT gene (0.43).



```
> fviz_silhouette(hc_cut_COI, main = "COI Silhouette")
  cluster size ave.sil.width
1       1   10          0.42
2       2    6          0.53
3       3    2          0.49
> # Visualize silhouhette information
> fviz_silhouette(hc_cut_CYTB, main = "CYTB Silhouette")
  cluster size ave.sil.width
1       1   10          0.41
2       2    6          0.41
3       3    2          0.56
```

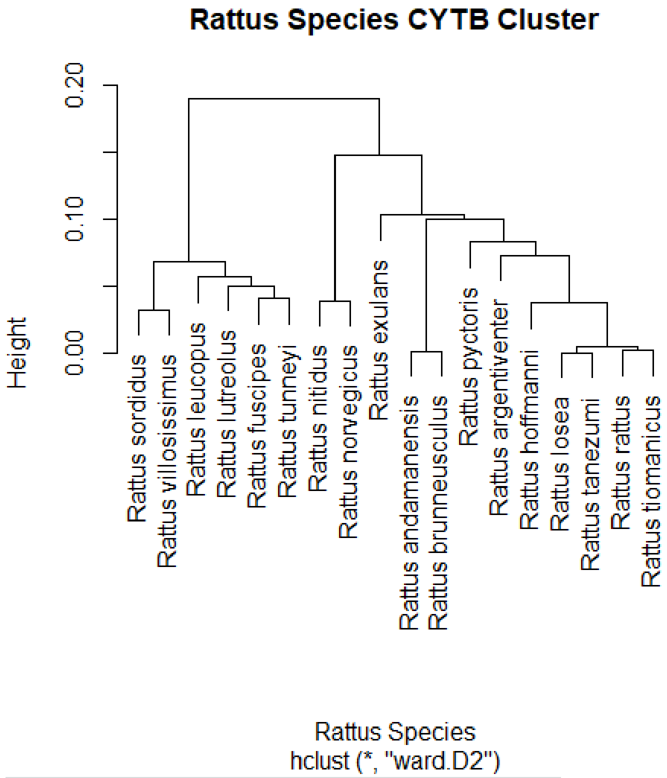
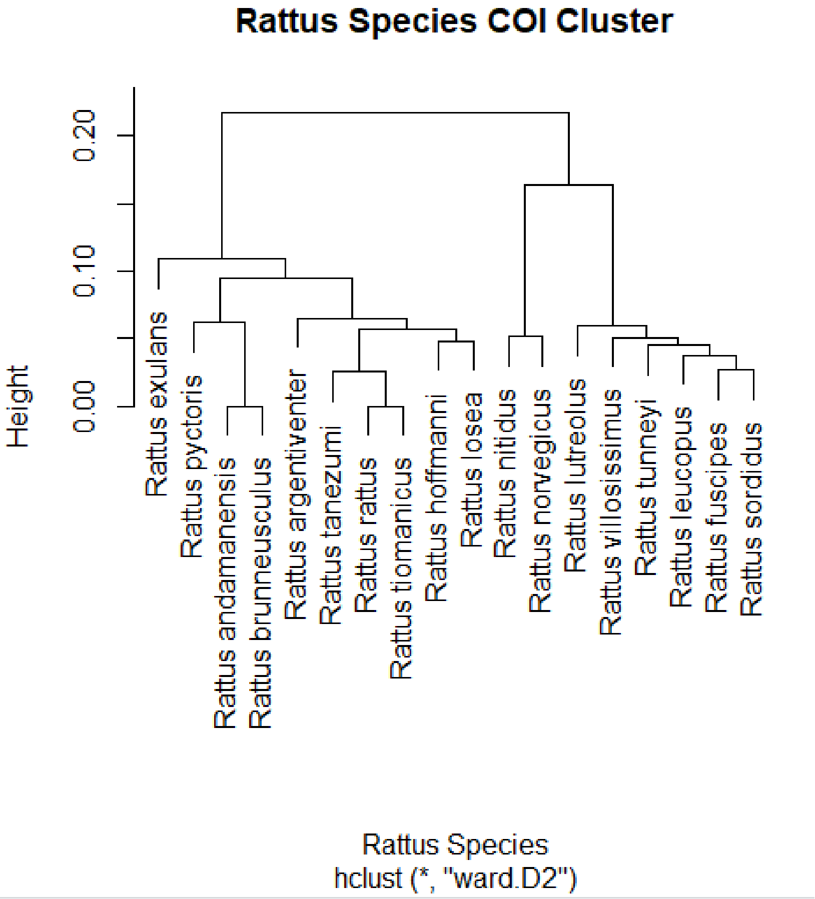


Figure 2. Phylogenetic tree of Rattus Species in accordance to CYTB and COI gene clusters



Results and Discussions

This experiment has indicated that COI and CYTB gene exhibit different clustering patterns within *Rattus* taxonomical group. Despite both being mitochondrial genes, there seems to be a huge difference in the phylogenetic tree of each respective gene cluster. The silhouette index scores however are rather low (0.46 COI, 0.43 for CYTB) which may indicate that the cluster strength isn't the best for either gene, and that maybe more accurate data is required for this process. One issue with the dataset was that only a few species had enough gene sequence data to not introduce biases in the alignment process. However, it was important to first align and find a consensus sequence amongst *Rattus* species before a final alignment and clustering process to get the wanted results. I believe going forward, a better set of data is necessary for any phylogenetic analysis. In addition, I believe I could have done a better job identifying and fixing the DNA sequences before analysis, and that a better understanding of clustering models and distance matrix models may have presented a better result.

References

- Baum, D. (2008) Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. Nature Education 1(1):190
- Yu, H., Jamieson, A., Hulme-Beaman, A., Conroy, C. J., Knight, B., Speller, C., Al-Jarah, H., Eager, H., Trinks, A., Adikari, G., Baron, H., Böhlendorf-Arslan, B., Bohingamuwa, W., Crowther, A., Cucchi, T., Esser, K., Fleisher, J., Gidney, L., Gladilina, E., Gol'din, P., … Orton, D. (2022). Palaeogenomic analysis of black rat (Rattus rattus) reveals multiple European introductions associated with human economic history. Nature communications, 13(1), 2399. https://doi.org/10.1038/s41467-022-30009-z
- Liu, Y., Yao, L., Ci, Y., Cao, X., Zhao, M., Li, Y., & Zhang, X. (2021). Genetic differentiation of geographic populations of Rattus tanezumi based on the mitochondrial Cytb gene. PloS one, 16(3), e0248102. https://doi.org/10.1371/journal.pone.0248102
- Yang, C. H., Wu, K. C., Chuang, L. Y., & Chang, H. W. (2019). Decision Theory-Based COI-SNP Tagging Approach for 126 Scombriformes Species Tagging. Frontiers in genetics, 10, 259. https://doi.org/10.3389/fgene.2019.00259

Acknowledgement

- **https://docs.ropensci.org/rentrez/ - Web history**
- **https://bradleyboehmke.github.io/HOML/hierarchical.html - Agnes clustering**
- I acknowledge Jessica Castellanos Labarcena for helping me develop the function to find the consensus sequence for each Rattus species. She was very helpful in guiding me through the code and understanding how to approach it.
- I acknowledge BINF6210 practice scripts which guided me on how to perform alignment, clustering, and data extraction.