The Phylogenetic Community Analysis of the *Odonata* Order in PEI National Park

Stephanie Rosen

12/12/2021

Introduction

Water pollution and climate change are two large ongoing issues that affect the health of many aquatic organisms across Canada. Water pollution due to biological contaminants or other inorganic toxins can pose threats to the health of an ecosystem. This project aims to use biodiversity as a marker of ecosystem health. There are many species that can used as indicator species in water ecosystems, especially macroinvertebrates (Dunbar et al., 2010). More specifically, *Odonata*, an order of dragonflies and damselflies, have been known to be especially useful for ecosystem monitoring (Hart et al., 2014).

Prince Edward Island National Park is located in Prince Edward Island and provides a conservation area for many plants and animals (Parks Canada). In order to maintain the health of the ecosystems in the park, it is important to investigate the levels of pollution of ecosystem health, especially in bodies of water. Since 2008, Parks Canada has been monitoring the abundance of *Odonata* of ponds within the National Park (Parks Canada). However, there is no further analysis publicly available about the diversity of the species and how the community structure might inform ecosystem health.

The objective of this study is to investigate the biodiversity of the order *Odonata* as a marker of ecosystem health in ponds of PEI National Park. This will be carried out by investigating phylogenetic community structure among three ponds. Using a phylogenetic tree visualization along with measures of diversity among communities, resulting in an exploration of biodiversity among *Odonata* species. This will assist in providing a better understanding of the health of each pond ecosystem.

Description of Dataset

The data investigated in this project was obtained from the Government of Canada on December 1, 2021 via a website download. This dataset was produced by *Odonata* monitoring from the years of 2008 to 2017 and provides species counts in ponds within the boundaries of the National Park. The main variables collected in this dataset include: date of collection, scientific name, common name, plot number, and species count. The species counts are produced from collections of larval exoskeletons each summer (June to August). The dataset includes 1,680 observations over 10 years.

Data Acquisition, Exploration, Filtering, and Quality Control

Here we are uploading our data from the Parks Canada database. The data was originally obtained on December 1st, 2021 from the below URL:

https://open.canada.ca/data/en/dataset/ce777149-3cb9-4ac2-a133-0924fef16c6e

```r
#Here we import the dataset from the CSV file
df_Odonata <- read.csv("PEI_NP_Wetlands_Odonata_2010-2018_data.csv", sep = ";")

#Looking at the variables and dimensions of the dataset
names(df_Odonata)
```

```
##  [1] "Scientific.name"        "Common.name"            "count"
##  [4] "Plot.number"            "Year"                   "Month"
##  [7] "Day"                    "Survey.site"            "exuvia.observer.number"
## [10] "Observer..order.number" "note"
```

```r
dim(df_Odonata)
```

```
## [1] 4365   11
```

```r
head(unique(df_Odonata$Scientific.name))
```

```
## [1] "Nom scientifique"         "Coenagrionidae sp."
## [3] "Lestes sp."               "Epitheca spinigera"
## [5] "Libellula quadrimaculata" "Leucorrhinia intacta"
```

```r
#It appears that there are lots of errors with species names in our dataset. With the below code, we wi

#Remove chunks of text in brackets but still have full species names
df_Odonata$Scientific.name <- str_remove(string = df_Odonata$Scientific.name, "\\(.+\\)" )

#There are some species that have a third unecessary word in their name and do not have proper capitali
df_Odonata$Scientific.name <- str_remove(string = df_Odonata$Scientific.name, "Walker" )
df_Odonata$Scientific.name <- str_remove(string = df_Odonata$Scientific.name, "Say" )
df_Odonata$Scientific.name <- str_replace(string = df_Odonata$Scientific.name, "anax", "Anax")
df_Odonata$Scientific.name <- str_replace(string = df_Odonata$Scientific.name, "leucorrhinia", "Leucorr
df_Odonata$Scientific.name <- str_replace(string = df_Odonata$Scientific.name, "aeshna", "Aeshna")

#We also remove the first row (French header), scientific names with NA values, irrelevant columns, and

df_Odonata <- df_Odonata[-1,] %>%
  filter(!is.na(Scientific.name)) %>%
  select(!exuvia.observer.number & !Observer..order.number & !note) %>%
  filter(str_detect(Scientific.name, "^[A-z]+ [A-z]+$")) %>%
  filter(!str_detect(Scientific.name, "NOT ODONATA")) %>%
  filter(!str_detect(Scientific.name, "unknown Aniaoptera")) %>% #remove observation with no species le
  filter(!str_detect(Scientific.name, "Enallagma sp")) #remove observation with no species level indica

#The names look much better now.
head(unique(df_Odonata$Scientific.name))
```

```
## [1] "Epitheca spinigera"       "Libellula quadrimaculata"
## [3] "Leucorrhinia intacta"     "Sympetrum obtrusum"
## [5] "Aeshna interrupta"        "Sympetrum costiferum"
```

```
#Now that we have clean scientific names, let's look at the different ponds we have.
unique(df_Odonata$Survey.site)
```

```
##  [1] "Bells Pond"         "Bog Pond"           "Bowley Pond"
##  [4] "John Archies Pond"  "Wills Pond"         "Barachois Pond"
##  [7] "Crowbush 7th tee"   "Crowbush 8th tee"   "Cumberland Pond"
## [10] "Harbour Road Pond"  "Nail Pond - West"   "Nicholson Rd. Pond"
## [13] "Pigots Pond"        "Rayner's Pond"      "Seaweed Rd. Pond"
```

```
#We have a tootal of 16 ponds. Let's see how many were monitored each year.

#Let's look at the number of counts per pond per year. We will group the observations by year and count
df_sites <-
  df_Odonata %>%
  group_by(Year) %>%
  count(Survey.site)

df_counts_by_site <-
  pivot_wider(df_sites, id_cols = Survey.site, names_from = Year, values_from = Year)

#From this table, we see that there were different sites measured every year. To keep the analysis cons

df_Odonata_common_sites <-
  df_Odonata %>%
  filter(Survey.site == "Bells Pond" | Survey.site == "Bog Pond" | Survey.site == "Bowley Pond")

#Now the df_Odonata_common sites data has all the data for the ponds measured every year. We are ready
```

Now we have clean scientific name data for our different communities. Let's explore the data.

```
#Exploring total number of counts at each pond
df_Odonata_counts_by_year_plot <- df_Odonata_common_sites %>%
  group_by(Survey.site) %>%
  count(Survey.site)

#Exploring total number of counts at each pond per year
df_Odonata_unique_by_year_pond <- df_Odonata_common_sites %>%
  group_by(Year, Survey.site) %>%
  summarise(n = n_distinct(Scientific.name))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
#We can see that Bwoley Pond has the most observations. Let's look at the number of observations per ye
ggplot(df_Odonata_unique_by_year_pond,
       aes(fill=Survey.site, y=n, x=Year)) +
  geom_bar(position="dodge", stat="identity") +
  labs(y = "Number of Unique Odonata Species",
       title = "Number of Unique Odonata Species per Pond by Year") +
  scale_fill_manual(values = c("#D81B60","#1E88E5","#FFC107"),
                    name = "Survey Site") +
  theme(plot.title = element_text(hjust = 0.5))
```

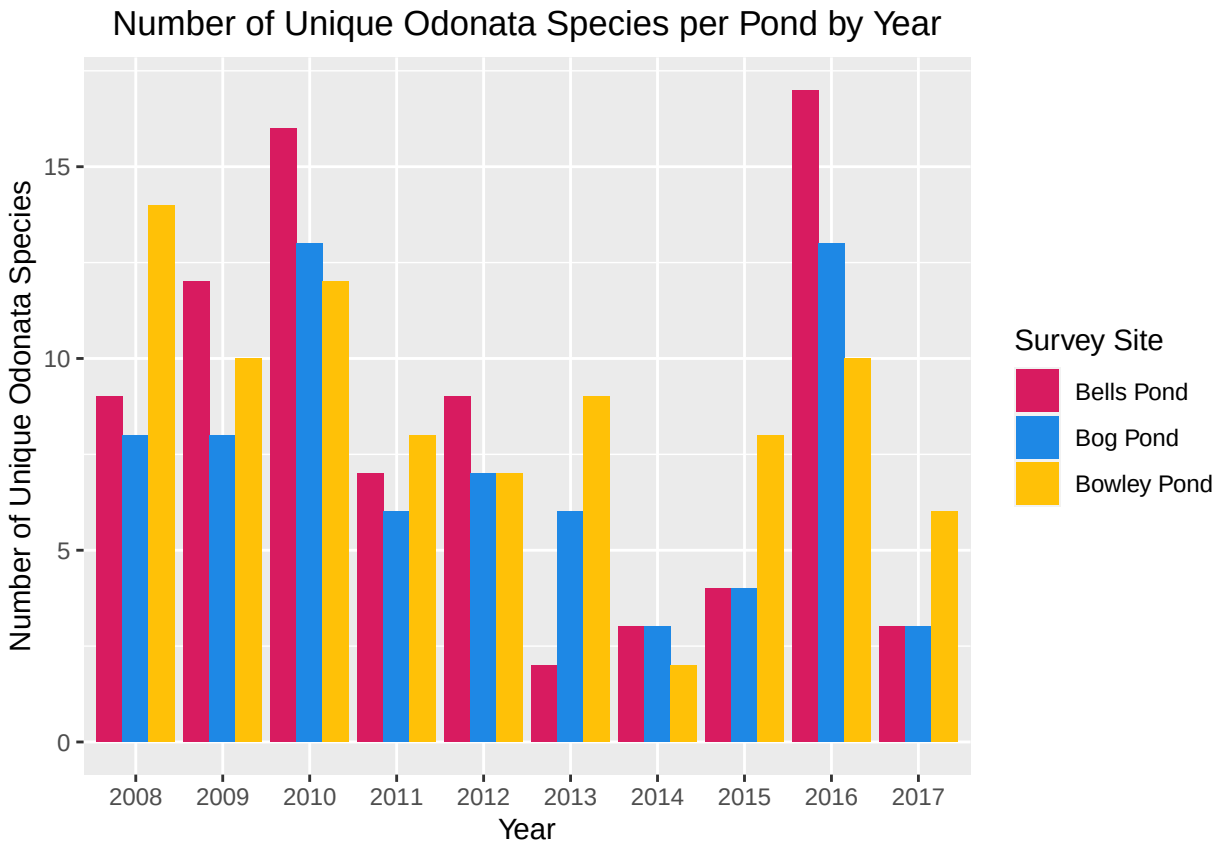

Figure 1: Unique number of *Odonata* members observed in PEI National Park ponds from 2008 to 2017. The ponds observed were Bells Pond, Bog Pond, and Bowley Pond.

```
#Now we can put our data into a community matrix format for future analysis with the picante package.

#The counts are currently in character format
class(df_Odonata_common_sites$count)
```

```
## [1] "character"
```

```
#Converting into numeric format so we can sum the counts for each species
df_Odonata_common_sites$count <- as.numeric(df_Odonata_common_sites$count)

#Putting into community data format using the tapply function to sum the counts for each species
df_Odonata_spread <- as.data.frame(tapply(X = df_Odonata_common_sites$count,
       INDEX = list(df_Odonata_common_sites$Survey.site,df_Odonata_common_sites$Scientific.name ), FUN =

#Replacing NA values with 0
df_Odonata_spread[is.na(df_Odonata_spread)] <- 0
```

Our community data is in the correct format now! Now we will need to access the NCBI

Now we will use the NCBI database to obtain our sequences. The first choice for sequences was 18S rRNA. Dumont et al. show that the 18S sequences for the order *Odonata* was able to resolve deep relationships and worked well for this taxonomic group.

```r
#Let's get a list of unique species grouped by ponds so we have a list of unique species to provide for

df_Odonata_unique_by_pond <- df_Odonata_common_sites %>%
  group_by(Survey.site, Scientific.name) %>%
  summarise()
```

```
## `summarise()` has grouped output by 'Survey.site'. You can override using the `.groups` argument.
```

```r
#Now we can get NCBI data for 18S sequences

eighteenS_search_ids <- vector() #Empty vector for all the search ids
unique_species <- unique(df_Odonata_common_sites$Scientific.name)
length(unique_species) #We have a total of 33 unique species
```

```
## [1] 33
```

```r
#We use a for loop to loop through and get a single ID for an 18S rRNA sequence for each species
for (species in unique_species) {
  eighteenS_search <- entrez_search(db = "nuccore", term = paste(species, "[ORGN] AND 18S rRNA AND biom
  eighteenS_search_ids <- append(eighteenS_search_ids, eighteenS_search$ids[1])
}

#If we take a look at this object, we see that it didn't return any hits at all! Even with adjusting th
head(eighteenS_search_ids)
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
```

```r
#When looking through the available sequences for some of the species in the NCBI database, most only h

COI_search_ids <- vector() #empty vector for all the search ids

#loop through and get the first ID for each species COI sequence
for (species in unique_species) {
  COI_search <- entrez_search(db = "nuccore", term = paste(species, "[ORGN] AND COI[GENE] AND 650:670[SL
  COI_search_ids <- append(COI_search_ids, COI_search$ids[1])
```

```
  }

#This looks much better!
length(COI_search_ids)
```

```
## [1] 33
```

```
head(COI_search_ids)
```

```
## [[1]]
## [1] "331127674"
##
## [[2]]
## [1] "2008116030"
##
## [[3]]
## [1] "821682846"
##
## [[4]]
## [1] "821680984"
##
## [[5]]
## [1] "821682832"
##
## [[6]]
## [1] "821660218"
```

```
#However, when we take a deeper look at the data, there are 4 species that have no sequence ID's. We wi

COI_search_ids <- COI_search_ids[-c(15, 20, 26, 29)]

#Now let's download the actual sequences from NCBI in fasta format.
COI_fetch <- entrez_fetch(db = "nuccore", id = COI_search_ids, rettype = "fasta")

#Writing  FASTA file to working directory
write(COI_fetch, "COI_fetch.fasta", sep = "\n")

#Converting the FASTA file to a BioStrings DNAStringSet format
stringSet <- readDNAStringSet("COI_fetch.fasta")
class(stringSet)
```

```
## [1] "DNAStringSet"
## attr(,"package")
## [1] "Biostrings"
```

```
head(names(stringSet))
```

```
## [1] "JF839398.1 Epitheca spinigera voucher 10-SKOD-269 cytochrome oxidase subunit 1 (COI) gene, parti
## [2] "MW490564.1 Libellula quadrimaculata voucher ODOPL_164 cytochrome oxidase subunit 1 (COI) gene, p
## [3] "KM537628.1 Leucorrhinia intacta voucher 08SOODO-0044 cytochrome oxidase subunit 1 (COI) gene, pa
## [4] "KM536697.1 Sympetrum obtrusum voucher 08BBODO-343 cytochrome oxidase subunit 1 (COI) gene, parti
## [5] "KM537621.1 Aeshna interrupta voucher 08OMSOD-0144 cytochrome oxidase subunit 1 (COI) gene, parti
## [6] "KM529127.1 Sympetrum costiferum voucher 08BBODO-259 cytochrome oxidase subunit 1 (COI) gene, pa
```

```r
#Put sequences into dataframe format
df_COI_Odonata <- data.frame(COI_Title = names(stringSet), COI_Sequence = paste(stringSet))
```

Now we have obtained all our COI sequences. Let's explore the sequence data.

```r
#explore the length of sequences using a boxplot:
seq_length <- str_count(df_COI_Odonata$COI_Sequence)
boxplot(seq_length, ylab = "COI Sequence Lengths (base pairs)", xlab = "COI Sequence")
```

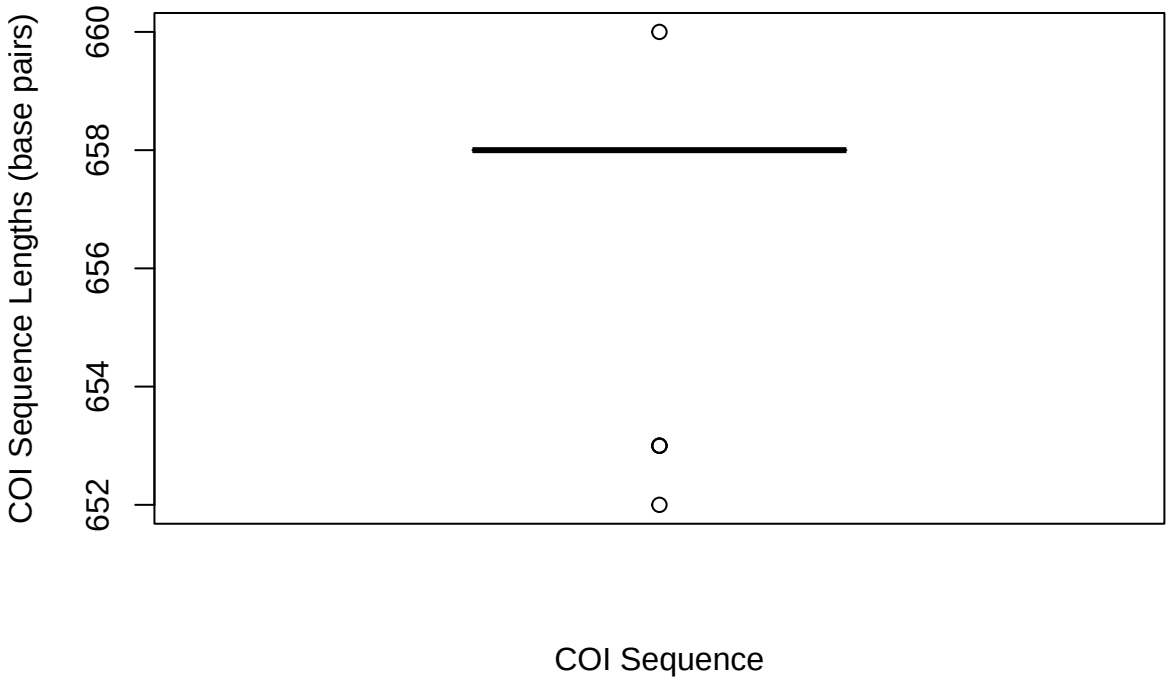

Figure 2: Cytochrome c oxidase subunit I (COI) sequences from species present in Parks Canada data prior to filtering. The COI sequences were retrieved from NCBI.

I have made the choice to include the outliers as shown by the boxplot in the analysis as they are within 10 base pairs of the median.

```r
#Since this is a shorter gene, I have only allowed for 1% of the data to be missing within the sequence
missing.data <- 0.01

#Since there is not a lot of natural variability in COI sequences, I have chosen a low number of allowe
length.var <- 10

#We will filter the data to remove sequences outside the allowed missing data and sequence length thres
df_COI_Odonata <- df_COI_Odonata %>%
  mutate(COI_Sequence2 = str_remove_all(COI_Sequence, "^N+|N+$|-")) %>%
  filter(str_count(COI_Sequence2, "N") <= (missing.data * str_count(COI_Sequence))) %>%
  filter(str_count(COI_Sequence2) >= median(str_count(COI_Sequence2)) - length.var & str_count(COI_Seque

#Cleaning up the species names to proper taxonomic names
df_COI_Odonata$Species_Name <- word(df_COI_Odonata$COI_Title, start = 2L, end = 3L)

remove(missing.data, length.var)
```

Main Software Tools Description

The main tool chosen for visualization was the ggtree package (Yu, 2020). A strength of this package is that it is similar in syntax to the ggplot2 package. This enabled a faster and more understandable coding experience. One weakness is that the tree had to be converted from a phylo format which provided some difficulty. The picante package was considered for the visualizations, however, it did not appear to have as many visualization options as the ggtreee package. I still chose to use the picante package for community structure analysis (Kembel et al., 2010). This is the main purpose of the picante software and had a well-written vigenette for beginners.

Main Analysis

A phylogenetic tree will be built and so we will first perform a sequence alignment.

```
df_COI_Odonata$COI_Sequence2 <- DNAStringSet(df_COI_Odonata$COI_Sequence2)

#Naming the sequences before they are aligned
names(df_COI_Odonata$COI_Sequence2) <- df_COI_Odonata$Species_Name

#aligning sequences using default settings
dfCOI.alignment <- DNAStringSet(muscle::muscle(df_COI_Odonata$COI_Sequence2, verbose = FALSE))
```

```
##
## MUSCLE v3.8.31 by Robert C. Edgar
##
## http://www.drive5.com/muscle
## This software is donated to the public domain.
## Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.
##
## file44ab786a7df6 29 seqs, max length 660, avg  length 657
## 543 MB(3%)00:00:00                Iter   1    0.23%  K-mer dist pass 1543 MB(3%)543 MB(3%)00:00:00
## 543 MB(3%)00:00:00                Iter   1    0.23%  K-mer dist pass 2543 MB(3%)543 MB(3%)00:00:00
## 545 MB(3%)00:00:00                Iter   1    3.57%  Align node       545 MB(3%)00:00:00
## 554 MB(3%)00:00:00                Iter   1    3.45%  Root alignment554 MB(3%)00:00:00
## 554 MB(3%)00:00:00                Iter   2    3.70%  Refine tree   555 MB(3%)00:00:00
## 555 MB(3%)00:00:00                Iter   2    3.45%  Root alignment555 MB(3%)00:00:00
## 555 MB(3%)555 MB(3%)00:00:00                Iter   2  100.00%  Root alignment
## 555 MB(3%)00:00:00                Iter   3    3.64%  Refine biparts555 MB(3%)00:00:00
```

```
#writing sequence alignment to file in FASTA format
writeXStringSet(dfCOI.alignment, file = "Odonata.fas", format = "fasta")

#look at sequences in browser
BrowseSeqs(dfCOI.alignment)

#This sequence alignment looks good! There are no obvious outliers.
```

Now, we will use the sequence alignment to build a phylogenetic tree. The clustering method chosen was maximum likelihood as suggested in a previous *Odonata* phylogenetic study (Kim et al., 2014). The nucleotide substitution model suggested by the same study was the general time reversible model (Kim et al., 2014). This model was not available in the utilized package and so the most similar option was chosen,

TN93. The TN93 model assumes distinct rates for both kinds of transition (A <-> G versus C <-> T), and transversions.

```
#Clustering!

chosen.model <- "TN93"
clustering.threshold <- 0.03
clustering.method <- "ML"

#Putting alignment into DNAbin format for the adjusted and non-adjusted alignments
dnaBin.COI <- as.DNAbin(dfCOI.alignment)

#Calculating the distance matrix
distanceMatrix <- dist.dna(dnaBin.COI, model = chosen.model, as.matrix = TRUE, pairwise.deletion = TRUE)


#Clustering using function IdClusters and our settings defined at the top for clustering method and clu
clusters.COI <- IdClusters(distanceMatrix,
                           method = clustering.method,
                           cutoff = clustering.threshold,
                         myXStringSet = dfCOI.alignment,
                           showPlot = FALSE,
                           type = "both",
                           verbose = TRUE)
```

```
## Constructing initial neighbor-joining tree:
## ============================================================================
##
## JC69:     -ln(L)=6444, AICc=13028, BIC=13192
## JC69+G4:  -ln(L)=5789, AICc=11722, BIC=11888
## K80:      -ln(L)=6228, AICc=12599, BIC=12766
## K80+G4:   -ln(L)=5647, AICc=11442, BIC=11610
## F81:      -ln(L)=6151, AICc=12454, BIC=12625
## F81+G4:   -ln(L)=5417, AICc=10988, BIC=11161
## HKY85:    -ln(L)=5881, AICc=11915, BIC=12088
## HKY85+G4: -ln(L)=5202, AICc=10562, BIC=10737
## T92:      -ln(L)=5890, AICc=11927, BIC=12095
## T92+G4:   -ln(L)=5213, AICc=10577, BIC=10748
## TN93:     -ln(L)=5830, AICc=11818, BIC=11993
## TN93+G4:  -ln(L)=5159, AICc=10480, BIC=10657
##
## The selected model was:  TN93+G4
##
## Maximizing Likelihood of Tree:
## -ln(Likelihood) = 5159 (0.00% improvement), 0 NNIs -ln(Likelihood) = 5148 (0.22% improvement), 1 NNI
##
## Model parameters:
## Frequency(A) = 0.305
## Frequency(C) = 0.171
## Frequency(G) = 0.172
## Frequency(T) = 0.352
## Rate A <-> G = 1.836
## Rate C <-> T = 4.453
## Transversion rates = 1
```

```
## Alpha = 0.339
##
## Time difference of 1.71 secs
```

```
#Create Tree
phyloCOI <- as.phylo.dendrogram(clusters.COI[[2]])


tree <-
  ggtree(phyloCOI,
         #branch.length="none",
       color = "black", size = 0.5) +
  theme_tree2() +
  ggtitle("Phylogenetic Tree of the Odonata Family in PEI National Park") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlim(NA, 0.24) +
  geom_tiplab(size=4, align=TRUE, linesize=.5, offset = 0.04)


species_to_remove <- c("Cordulia shurtleffi","Sympetrum rubicundulum", "Sympetrum obstrusum", "Celithem

df_Odonata_spread2 <- as.data.frame(tapply(X = df_Odonata_common_sites$count,
       INDEX = list(df_Odonata_common_sites$Scientific.name,df_Odonata_common_sites$Survey.site), FUN =

df_Odonata_spread2 <- df_Odonata_spread2[!(row.names(df_Odonata_spread2) %in% species_to_remove),]

df_Odonata_spread2[is.na(df_Odonata_spread2)] <- 0


#The below for loop will check which ponds each species is present in and add it to a variable called P
df_counts_for_tree <- data.frame(Species = row.names(df_Odonata_spread2), Pond = "")

for (species in row.names(df_Odonata_spread2)){

  #if species is in all ponds
  if (df_Odonata_spread2[species,]$`Bells Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` > 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "All"
  }

  #if species is in Bells and Bog
  else if(df_Odonata_spread2[species,]$`Bells Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` == 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bells and Bog"
  }

  #if species is in Bowley and Bog
  else if(df_Odonata_spread2[species,]$`Bells Pond` == 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` > 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bowley and Bog"
  }
```

```
  #if species in Bells and Bowley
  else if (df_Odonata_spread2[species,]$`Bells Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` == 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` > 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bells and Bowley"
  }

  #if species in Bells
  else if (df_Odonata_spread2[species,]$`Bells Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` == 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` == 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bells"
  }

  #if species in Bog
  else if (df_Odonata_spread2[species,]$`Bells Pond` == 0 &&
      df_Odonata_spread2[species,]$`Bog Pond` > 0 &&
      df_Odonata_spread2[species,]$`Bowley Pond` == 0) {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bog"
  }

  #if species in Bowley
  else {
    df_counts_for_tree[df_counts_for_tree$Species == species,]$Pond <- "Bowley"
  }

}


tree <- tree %<+% df_counts_for_tree #adding the pond data to the tree so we can include pond occurence

  tree2 <- tree +
  geom_tippoint(aes(colour = Pond, shape = Pond, fill = Pond), size = 4) + #Add points to the tips of t
  scale_shape_manual("Pond",
                     values = c("Bells" = 4,
                                "Bowley" = 10,
                                "Bog" = 15,
                                "Bowley and Bog" = 16,
                                "Bells and Bowley" = 24,
                                "Bells and Bog" = 18,
                                "All" = 0),
                     labels = c("Bells", "Bowley", "Bog", "Bowley and Bog", "Bells and Bowley", "Bells a

  scale_color_manual("Pond",
                     values = c("Bells" = "blue",
                                "Bowley" = "red",
                                "Bog" = "green",
                                "Bowley and Bog" = "yellow",
                                "Bells and Bowley" = "orange",
                                "Bells and Bog" = "black",
                                "All" = "brown"),
                     labels = c("Bells", "Bowley", "Bog", "Bowley and Bog", "Bells and Bowley", "Bells a
```

```
  scale_fill_manual("Pond",
                    values = c("Bells" = "blue",
                               "Bowley" = "red",
                               "Bog" = "green",
                               "Bowley and Bog" = "yellow",
                               "Bells and Bowley" = "orange",
                               "Bells and Bog" = "black",
                               "All" = "brown"),
                    labels = c("Bells", "Bowley", "Bog", "Bowley and Bog", "Bells and Bowley", "Bells a
  theme(legend.position = "right")  +
  scale_size_continuous(range = c(3, 10))
```

```
## Warning: Removed 1 rows containing missing values (geom_point_g_gtree).
```

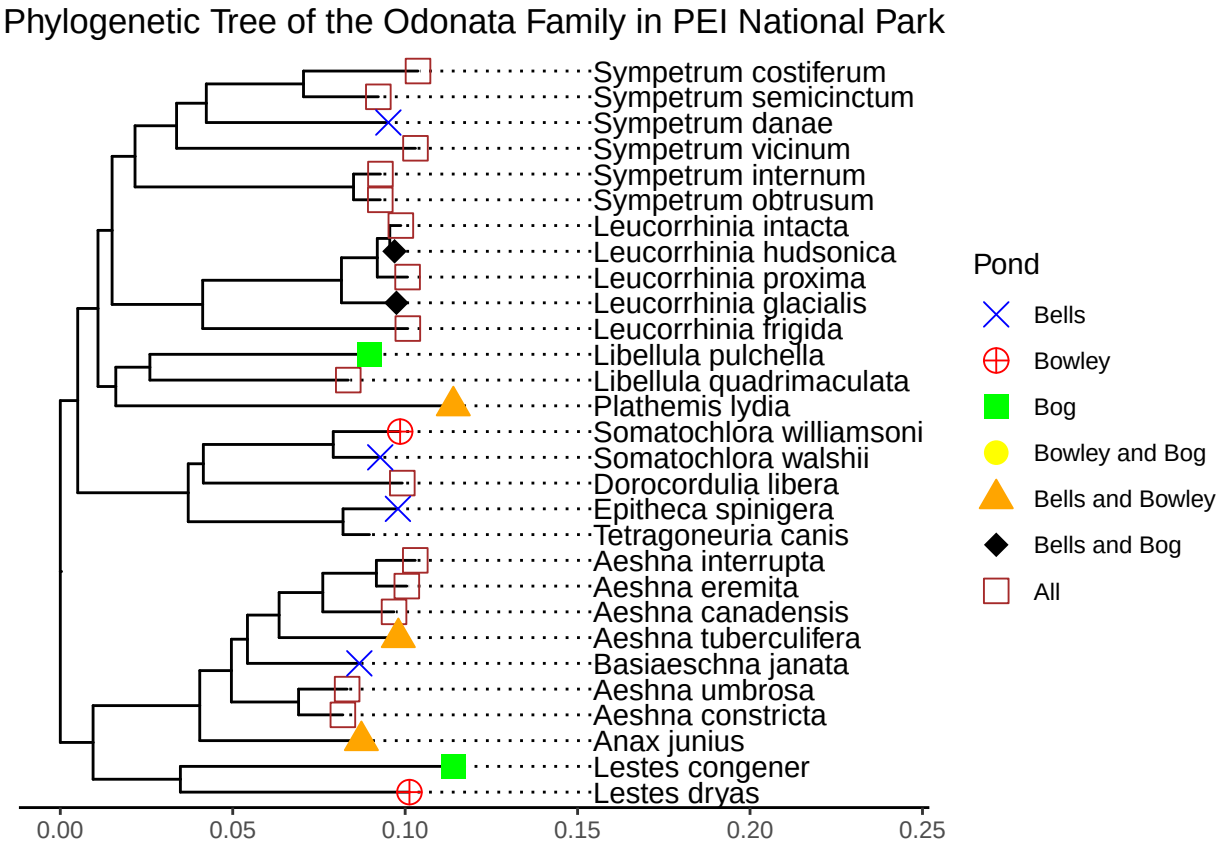

Figure 4: Phylogenetic tree using maximum likelihood methods and a TN93 clustering approach with pond location on the tips of the tree. Colour and shape of each branch tip represent the pond that each species has been observed at.

Now that we have a visualization of the data across communities, let's look at a diversity measure to quantify it within the picante package.

```
#making sure the tips of the tree match the species in our community data
cleanTree <- drop.tip(phy = phyloCOI,
                                  tip = setdiff(phyloCOI$tip.label,
                                  colnames(df_Odonata_spread)))
```

```
Odonata.PD <- pd(samp = df_Odonata_spread,
                 tree = cleanTree,
                 include.root = FALSE)

head(Odonata.PD)
```

```
##                    PD SR
## Bells Pond  1.1651516 27
## Bog Pond    0.9565197 22
## Bowley Pond 1.0992956 23
```

```
richness_diversity_plot<- plot(Odonata.PD$PD,
                               Odonata.PD$SR,
                               xlab = "Phylogenetic Diversity", ylab = "Species Richness",
                               pch = 16) + text(Odonata.PD$PD,Odonata.PD$SR,
     labels = row.names(Odonata.PD),
     cex = 0.8, pos = c(2,4,4), col = "dark green")
```

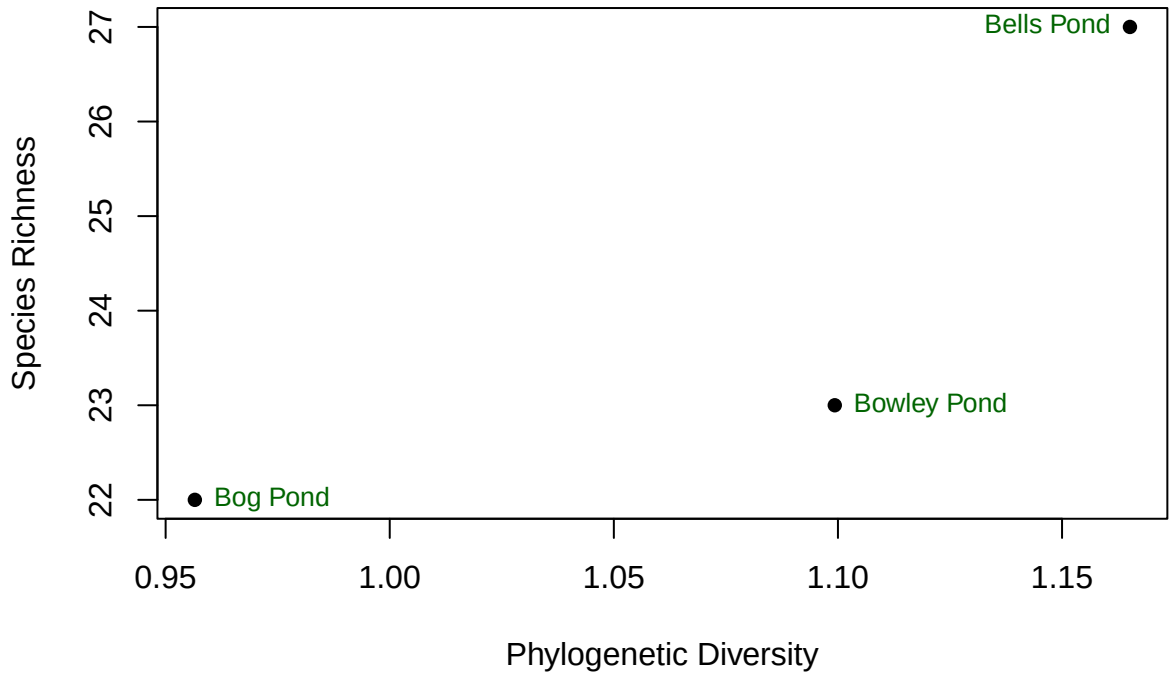

Figure 3: Phylogenetic diversity as a function of species richness for the *Odonata* order in Bog Pond, Bowley Pond, and Bells Pond in PEI National Park.

Results and Discussion

The main objective of the analysis was to investigate biodiversity of the *Odonata* order in three ponds of PEI National Park. This was carried out mainly by constructing a phylogenetic tree. In Figure 4, it can be shown that many of the species were present in each of the three ponds (Bowley, Bog, and Bell). It appears that these communities are overall quite similar with respect to the *Odonata* family. There are, however, four, three, and two species only found in Bells Pond, Bowley Pond, and Bog Pond respectively. This indicates

that there are some notable differences in the species diversity. No significant conclusions can be made with regard to specific branches being more present in certain ponds than others.

According to Faith & Baker, phylogenetic diversity (PD) values are defined as the minimum total length of all the phylogenetic branches required to span a given set of taxa on the phylogenetic tree (Faith & Baker, 2006). This means that larger PD values are correlated with greater diversity. Lower PD values can suggest that there is less diversity in a specific region and can indicate higher conservation priority. In this case, we see in Figure 5, that Bog Pond has the lowest PD value.

Though there are meaningful results from this analysis, it is not without its caveats. A point that is important to mention is that this study was completed using total species counts over a ten-year period. This is a very long period of time and does not acknowledge that biodiversity of *Odonata* is very likely to change between each year. In addition, there were many ponds that were excluded from analysis because they were not monitored each year. For consistency, they were excluded from the analysis, however, they hold lots of additional data that would be beneficial for future analysis. The quality of the dataset also presented an area of bias. There were many species that were only identified at the genus level. The phylogenetic tree and diversity analysis were completed at the species level. Therefore, any observations with inadequate species notation had to be excluded. A final consequence of this workflow is that the pond size was not taken into account. The species diversity revealed in each pond could be partially contributed to the size of its pond.

As stated in the introduction, macro-invertebrates are often used at indicator species. Within this analysis, only the order *Odonata* was analyzed. For future analysis, it would be interested to combine the results of this project with phylogenetic community analysis of other orders or families of macro-invertebrates. With regards to the lack of sequence data for 18S rRNA in NCBI, it may be worthwhile to explore alternate databases. Using the 18S sequences for this analysis would also provide an interesting comparison to the COI phylogenetic tree. With this dataset, there were also counts of each species per year per pond. As we were only interested in looking at the overall biodiversity, this data was not included. This set of data represents an opportunity to analyze species abundance in addition to biodiversity. On this note, it would be useful to also look at species abundance and tolerance to certain environmental factors. If a certain member of the *Odonata* order is more abundant than others, this species could very well be more tolerant to pollution than others.

Reflection

It's really amazing to see how far I've come! I think the most important thing that I have learned from this semester and this project is that I will never stop learning. There will always be new functions, packages, and applications to investigate and test out. I have also realized how time consuming the data exploration, quality control, and filtering can be. This step of the project took me more time than the main analysis. Going forward, I hope to solidify my data wrangling skills and become much more comfortable with manipulating data frames. I still find myself searching for data manipulation functions which can be quite time-consuming. I have also realized the benefits of project management and reducing large projects into manageable pieces. I will definitely take this forward in coursework and future academic projects.

Acknowledgements

Thank you to Jacqueline and Sally for the numerous example scripts and in-class demonstrations that have contributed to this project.

References

Community phylogenetics in R - github pages. (n.d.). Retrieved December 17, 2021, from https://pedrohbraga.github.io/CommunityPhylogenetics-Workshop/CommunityPhylogenetics-Workshop.html

Dumont, H. J., Vierstraete„ A., & Vanfleteren, J. R. (2010). A molecular phylogeny of the Odonata (Insecta). *Systematic Entomology*, 35(1), 6–18. Retrieved from https://doi.org/10.1111/j.1365-3113.2009.00489.x

Dunbar, M. J., Warren, M., Extence, C., Baker, L., Cadman, D., Mould, D. J., … Chadd, R. (2010). Interaction between macroinvertebrates, discharge and physical habitat in upland rivers. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 20(S1), S31–S44. Retrieved from https://doi.org/10.1002/aqc.1089

Explain GGPLOT2 warning: "removed K rows containing missing values". Stack Overflow. Retrieved December 17, 2021, from https://stackoverflow.com/questions/32505298/explain-ggplot2-warning-removed-k-rows-containing-missing-values

Faith, D. P., & Baker, A. M. (2006). Phylogenetic Diversity (PD) and Biodiversity Conservation: Some Bioinformatics Challenges. *Evolutionary Bioinformatics*, 2, 117693430600200000. Retrieved from https://doi.org/10.1177/117693430600200007

Hart, L. A., Bowker, M. B., Tarboton, W., & Downs, C. T. (2014). Species Composition, Distribution and Habitat Types of Odonata in the iSimangaliso Wetland Park, KwaZulu-Natal, South Africa and the Associated Conservation Implications. *PLoS ONE*, 9(3), e92588. Retrieved from https://doi.org/10.1371/journal.pone.0092588

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., … Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics (Oxford, England)*, 26(11), 1463–4. Retrieved from https://doi.org/10.1093/bioinformatics/btq166

Kim, M. J., Jung, K. S., Park, N. S., Wan, X., Kim, K., Jun, J., … Kim, I. (2014). Molecular phylogeny of the higher taxa of Odonata (Insecta) inferred from COI, 16S rRNA, 28S rRNA, and EF1-alpha sequences. *Entomological Research*, 44(2), 65–79. Retrieved from https://doi.org/10.1111/1748-5967.12051

Parks Canada Agency, G. of C. (2021, October 15). *Prince Edward Island national park*. Prince Edward Island National Park. Retrieved December 17, 2021, from https://www.pc.gc.ca/en/pn-np/pe/pei-ipe

Parks Canada. (n.d.). *Odonata - Prince Edward Island*. Open Data Canada. Retrieved December 17, 2021, from https://open.canada.ca/data/en/dataset/ce777149-3cb9-4ac2-a133-0924fef16c6e

Picante intro - cran.r-project.org. (n.d.). Retrieved December 17, 2021, from https://cran.r-project.org/web/packages/picante/vignettes/picante-intro.pdf

"plot.new has not been called yet" error in rmarkdown (Rstudio 1.0.44). Stack Overflow. (1965, January 1). Retrieved December 17, 2021, from https://stackoverflow.com/questions/40938561/plot-new-has-not-been-called-yet-error-in-rmarkdown-rstudio-1-0-44

Yu, G. (n.d.). *Ggtree: Elegant graphics for phylogenetic tree visualization and annotation*. Guangchuang Yu. Retrieved December 17, 2021, from https://guangchuangyu.github.io/ggtree-book/chapter-ggtree.html

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, 69(1), e96. Retrieved from https://doi.org/10.1002/cpbi.96