Exploratory Study of *Ixodes* Phylogeny and Species Distribution in Canada

Cynthia (Xiao Lei) Du

Department of Bioinformatics, University of Guelph

BINF 6210: Software Tools for Biological Data Analysis and Organization

Dr. Sarah (Sally) Adamowicz

December 13, 2023

Github Link: github.com/xdu006/6210A5-Geophylogeny

Introduction

As global warming continues, species previously confined to warmer regions due to climate limitations are migrating towards previously colder areas. *Ixodes spp.* (commonly named ticks) are a major taxon of concern as the warming trend continues due to their potential to carry a variety of tick-borne viral, bacterial, and protozoan diseases in urban areas (Rizzoli et al., 2014). One such disease, Lyme disease, is transferred by *Ixodes spp.* carrying the bacteria *Borrelia burgdorferi*. Lyme disease gathers increasing interest due to the difficulty of diagnosis, potentially serious symptoms, and lingering cognitive and neurological changes post-treatment (Marvel et al., 2022). In particular, both Europe and North America have experienced increased incidences of Lyme disease as climates warm (Cayol et al., 2018, Eisen & Eisen, 2018). In Canada, existing studies have been performed to determine the geographical distribution of *Ixodes scapularis*, the main carrier of *Borrelia burgdorferi*, in select Ontario regions (Slatculescu et al., 2020). However, no studies have covered the entirety of Canada and apprised the geophylogenic relationship between different *Ixodes spp.* Although few incidences of *Borrelia burgdorferi* have been detected in *Ixodes spp.* outside of *Ixodes scapularis*, closely related or variant species of *Ixodes* may become able to host the same pathogen due to genetic similarities. Additionally, as *Borrelia burgdorferi* can be transferred between ticks through consumption of the same infected food source, close proximity and resource sharing between sister species may allow the bacterium to adapt to and infect a new host. With global warming expanding the range of *Ixodes spp.*, it becomes important to conduct geophylogeny analysis to determine the proximity of closely related *Ixodes spp.* and understand the risk of zoonotic spillover to sister species. Overall, this exploratory report aims to determine the proximity of closely related *Ixodes*

spp. in Canada and aims to evaluate the risk of a new *Ixodes* host for *Borrelia burgdorferi* due to phylogenetic and geographic nearness.

Description of Data Set

The dataset used for this assignment was sourced from BOLD Systems (Hebert, 2013) through searching for taxons matching "Ixodes" and countries matching "Canada" on Dec 13, 2023. The dataset contained a total of 480 samples, including data for COI sequences, species names, latitude, and longitude, and was reduced to 337 samples following filtering and cleaning.

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

```
###IMPORT LIBRARIES

library(tidyverse)
library(dplyr)
library(ggforce)
library(ape)
library(phangorn)
library(phytools)
library(mapdata)
library(viridis)
library(DECIPHER)
library(Biostrings)

###DEFINING FUNCTIONS

#Function to output a sequence length distribution histogram given a set of
seqeunces
sequenceLengthDistributionHistogram <- function(nucleotidesColumn) {
  hist(nchar(nucleotidesColumn),
       main = "Sequence Length Distributions",
       xlab = "Sequence Length",
       ylab = "Number of Sequences") }

###AQUIRE AVAILABLE DATA

#Get data from BOLD
dfIxodesFull <- read_tsv(file =
("http://www.boldsystems.org/index.php/API_Public/combined?taxon=Ixodes&geo=C
anada&format=tsv"))
```

```r
#Check Dataset
#summary(dfIxodesFull)
table(dfIxodesFull$species_name)
```

```
##
##    Ixodes angustus      Ixodes banksi      Ixodes cookei    Ixodes
gregsoni
##                  3                  2                  2
2
##       Ixodes kingi       Ixodes marxi       Ixodes muris   Ixodes
pacificus
##                  1                  2                  2
1
##     Ixodes rugosus  Ixodes scapularis     Ixodes soricis Ixodes
spinipalpis
##                  2                436                  9
6
##     Ixodes texanus       Ixodes uriae
##                  1                  1
```

```r
#Remove all entries with missing species names, locations, or sequences, and
ensure that sequences are all COI-5P
dfIxodesDataAvailable <- subset(dfIxodesFull, is.na(species_name) == F &
                                is.na(lat) == F &
                                is.na(lon) == F &
                                is.na(nucleotides) == F &
                                grepl("COI-5P", marker_codes) == T)

#Check difference in number of total species.
print(sprintf("Number of species prior to cleaning: %s",
length(unique(dfIxodesFull$species_name))))
```

```
## [1] "Number of species prior to cleaning: 15"
```

```r
print(sprintf("Number of species after cleaning: %s",
length(unique(dfIxodesDataAvailable$species_name))))
```

```
## [1] "Number of species after cleaning: 14"
```

```r
print(sprintf("Total BINs reduced: %s",
length(unique(dfIxodesFull$species_name))-length(unique(dfIxodesDataAvailable
$species_name))))
```

```
## [1] "Total BINs reduced: 1"
```

```r
#Check difference in number of total number of records.
print(sprintf("Number of records prior to cleaning: %s", nrow(dfIxodesFull)))
```

```
## [1] "Number of records prior to cleaning: 480"
```

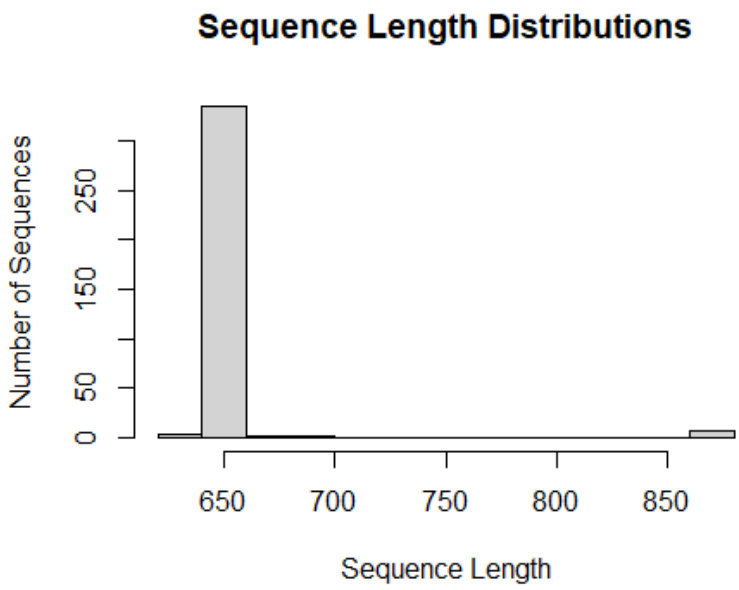
```r
print(sprintf("Number of records after cleaning: %s",
nrow(dfIxodesDataAvailable)))
```

```
## [1] "Number of records after cleaning: 346"
```

```
print(sprintf("Total records reduced: %s",
nrow(dfIxodesFull)-nrow(dfIxodesDataAvailable)))
```

```
## [1] "Total records reduced: 134"
```

```
#Note: record reduced are withing acceptable ranges
```




```
###VISUALIZATION 1

#Create Graphic for Ixodes Species Distribution in Canada
SpeciesDistributionGraph <-
ggplot(as.data.frame(table(dfIxodesDataAvailable$species_name)[order(table(df
IxodesDataAvailable$species_name))]), aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity") +
  facet_zoom(ylim = c(0, 10)) +
  geom_text(aes(label = Freq), vjust = -0.5, col="#3A3B3C") +
  ggtitle("Ixodes Reported Species Distribution in Canada") +
  xlab("Species")+
  ylab("Frequency Reported on BOLD") +
  guides(fill=guide_legend(title="Species")) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.text.y=element_text(size = 10),
        axis.title.x=element_text(size = 12, face="bold"),
        axis.title.y=element_text(size = 12, face="bold"),
        plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
        text = element_text(family = "AppleGothic"))
SpeciesDistributionGraph
```

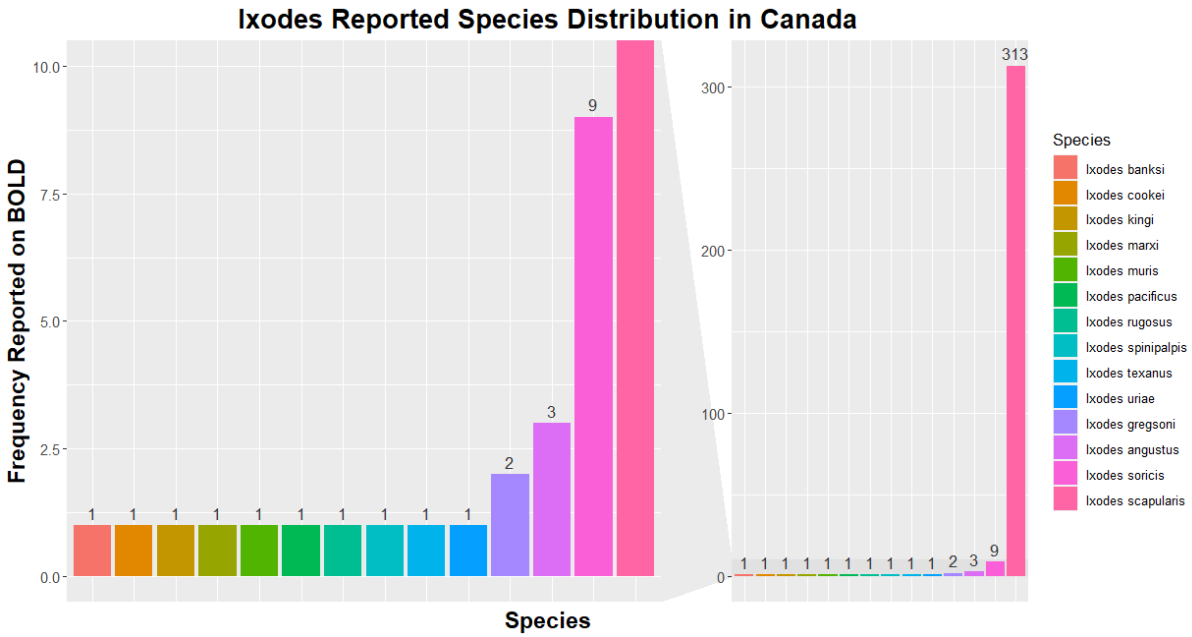
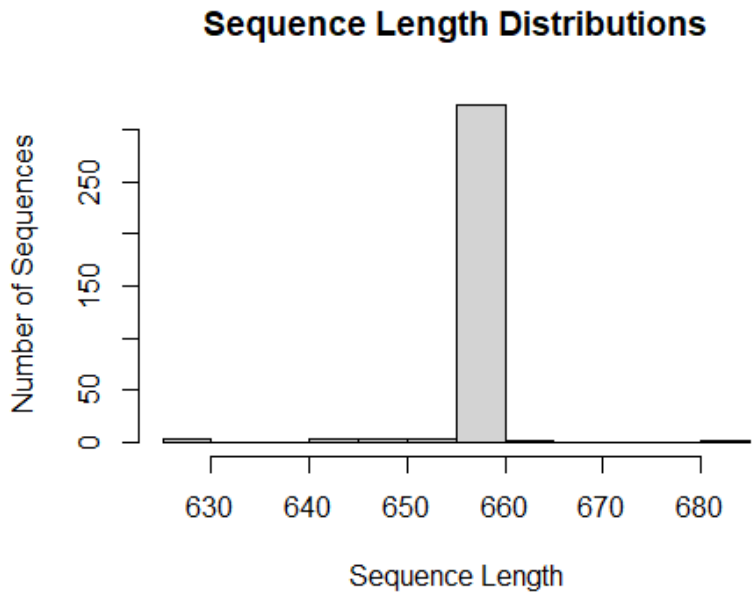


Figure 1. Distribution of BOLD entries available for each species of *Ixodes* in Canada. Miniture graph to the left represents a zoomed in view of the same data. Record counts are labeled as numbers above respective bars. Majority species is *Ixodes scapularis* as expected due to concern as main carrier of lyme disease.



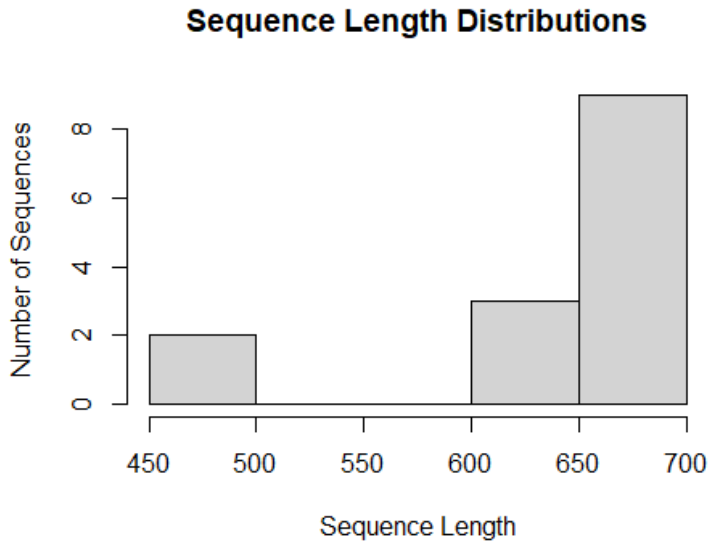
Main Software Tools Description

All processing for this assignment was completed in R. In terms of alignment, we used MUSCLE (Edgar, 2004), a widely used alignment algorithm. Construction of phylogenies was attempted with DECIPHER clustering (Wright, 2016), Maximum Likelihood (with neighbor-joining), and the highest-ranked model from ModelTest (both from the phangorn library; Schllep, 2011). Final geophylogeny plotting was performed only with DECIPHER cluster data and completed with phytools (Revell, 2012). Preferentially, we would have used a more powerful algorithm such as Maximum Likelihood or ModelTest that is able to model non-uniform evolutionary time, but we were unable to convert the output datatypes to the ones required for the mapping graphic. Instead, we simply use these additional phylogenies for comparison after initial analysis of species distribution on our map graphic. Consequently, one major weakness of the phytools package for geophylogeny mapping is the poor support for other common phylogeny data types such as pml and non-uniform length phylogenies.

Code Section 2 – Main Analysis

```
###PREPARE DATA FOR ALIGNMENT
#Clean Seqeunce Data and Pick Random Representitive Sequences for Each
Species

#Set up some parameters for sequence structure & alignment
missing.data <- 0.01
length.var <- 50
chosen.model <- "TN93"
clustering.threshold <- 0.03
clustering.method <- "single" #single linkage

#Clean sequences
dfIxodesDataAvailable <- dfIxodesDataAvailable %>%
  mutate(nucleotides2 = str_remove_all(nucleotides, "^N+|N+$|-")) %>%
#Remove starting and ending gaps or Ns
  filter(str_count(nucleotides2, "N") <= (missing.data *
str_count(nucleotides)))  #Remove sequences with greater than 1% Ns
  #length is not filtered as we did the histogram analysis earlier already
  #and the only representation of some species may be filtered out
```

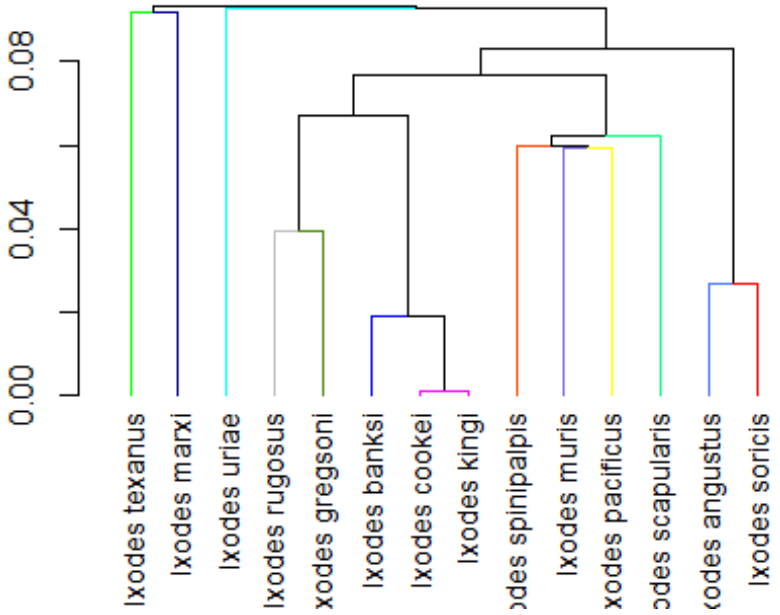
```r
#Pick one sequence to represent each species

#prime dataframe with some data
#for every species in the list of species names in available data,
#get a subset of of that species and randomly choose one row
#finally, remove the primer
set.seed(200)
repSequences <- dfIxodesDataAvailable[1,]
for(species in unique(dfIxodesDataAvailable$species_name)) {
  repSequences <- add_row(repSequences, subset(dfIxodesDataAvailable,
species_name == species)[sample(nrow(subset(dfIxodesDataAvailable,
species_name == species)), 1), ] ) }
repSequences <- repSequences[-1,]

#Check length distribution is appropriate
sequenceLengthDistributionHistogram(repSequences$nucleotides2)
#Note: good length distribution
```




```r
###CREAT DISTANCE MATRIX

#Change to appropriate data format for alignment
repSequences <- as.data.frame(repSequences) #change to dataframe
repSequences$nucleotides3 <- DNAStringSet(repSequences$nucleotides2) #Add
column with DNA stringset
names(repSequences$nucleotides3) <- repSequences$species_name #Add names to
DNA stringset
repSequences$nucleotides3 #Preliminary examination of sequences

## DNAStringSet object of length 14:
##      width seq                                               names
##  [1]   640 AACTATATACTTAATTTTTGGGA...AGCAGGTGGGGGAGATCCAATT Ixodes soricis
##  [2]   656 CTATATACTTAATTTTTGGGAGT...GATTCTATATCAACATTTATTC Ixodes
```

```
angustus
##  [3]   658 AACTATATATTTAATTTTTGGAA...TATTTTATATCAACATTTATTT Ixodes
scapularis
##  [4]   493 TTTTTTTATAGTTATACCAGTAA...AGGGGGTGGAGACCCTATTCTT Ixodes kingi
##  [5]   620 TTTGGAAGATGGTCAACTATAGT...ACCCCTCAGGGGGTGGAGACCC Ixodes cookei
##  ...   ... ...
## [10]   658 AACTATATATTTAATTTTTGGGA...TATTTTATACCAACATTTATTC Ixodes
pacificus
## [11]   603 AACTATATATTTAATTTTTGGGA...TAACAGACCGAAATTTTAATAC Ixodes muris
## [12]   658 AACTATATACTTAATTTTTGGGA...TATTCTTTATCAACATTTATTT Ixodes rugosus
## [13]   474 AACTATATATTTAATTTTTGGTA...CAATTATCAATATACGTTCCCC Ixodes
spinipalpis
## [14]   658 AACAATATATTTAATTTTTGGGA...TATTCTATACCAACACTTATTT Ixodes texanus

BrowseSeqs(repSequences$nucleotides3) #Preliminary examination of sequences

#Align sequences
alignment <- DNAStringSet(muscle::muscle(repSequences$nucleotides3, gapopen =
-1000, log = "log.tx", verbose = T), use.names = T)

alignment #view alignment

## DNAStringSet object of length 14:
##      width seq                                               names
##  [1]   659 AACTATATACTTAATTTTTGGGA...ATT------------------- Ixodes soricis
##  [2]   659 --CTATATACTTAATTTTTGGGA...ATTCTATATCAACATTTATTC- Ixodes
angustus
##  [3]   659 AACTATATATTTAATTTTTGGAA...ATTTTATATCAACATTTATTT- Ixodes
scapularis
##  [4]   659 ------------------------...ATTCTT---------------- Ixodes kingi
##  [5]   659 ----------------TTTGGAA...---------------------- Ixodes cookei
##  ...   ... ...
## [10]   659 AACTATATATTTAATTTTTGGGA...ATTTTATACCAACATTTATTC- Ixodes
pacificus
## [11]   659 AACTATATATTTAATTTTTGGGA...---------------------- Ixodes muris
## [12]   659 AACTATATACTTAATTTTTGGGA...ATTCTTTATCAACATTTATTT- Ixodes rugosus
## [13]   659 AACTATATATTTAATTTTTGGTA...---------------------- Ixodes
spinipalpis
## [14]   659 AACAATATATTTAATTTTTGGGA...ATTCTATACCAACACTTATTT- Ixodes texanus

BrowseSeqs(alignment) #view alignment in browser
#tested a couple other conditions, this one was great

#convert to DNAbin
alignmentBIN <- as.DNAbin(alignment)

#create distance matrix between sequences
distanceMatrix <- dist.dna(alignmentBIN, model = chosen.model, as.matrix =
TRUE, pairwise.deletion = TRUE)
```

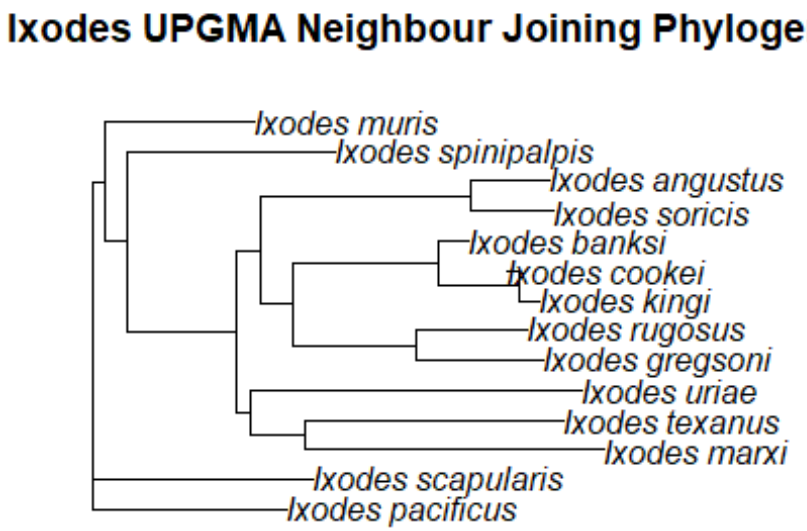
```
###CONSTRUCT PHYLOGENY

#clustering (view plot)
clusters <- DECIPHER::TreeLine(myDistMatrix = distanceMatrix,
                               method = clustering.method,
                               cutoff = clustering.threshold,
                               showPlot = TRUE,
                               type = "dendrogram",
                               verbose = TRUE)
```




```
clusters
```

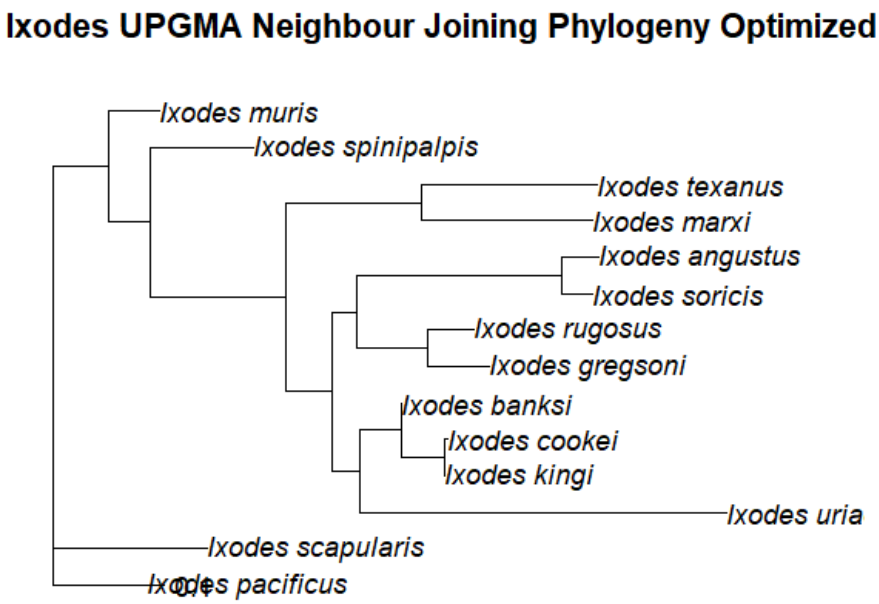
```
## 'dendrogram' with 2 branches and 14 members total, at height 0.09335777
```

```
#convert datatype for plotting
phyloData <- as.hclust(clusters)
phyloData <- as.phylo(phyloData)

###MAXIMUM LIKELIHOOD PHYLOGENY (NOT USED IN GRAPH)
#Other phylogeny methods, but failed to get correct data format for plotting

#try maximum likelihood

#make neighbour joining tree
njTree <- nj(distanceMatrix)
plot(njTree, main="Ixodes UPGMA Neighbour Joining Phylogeny")
```

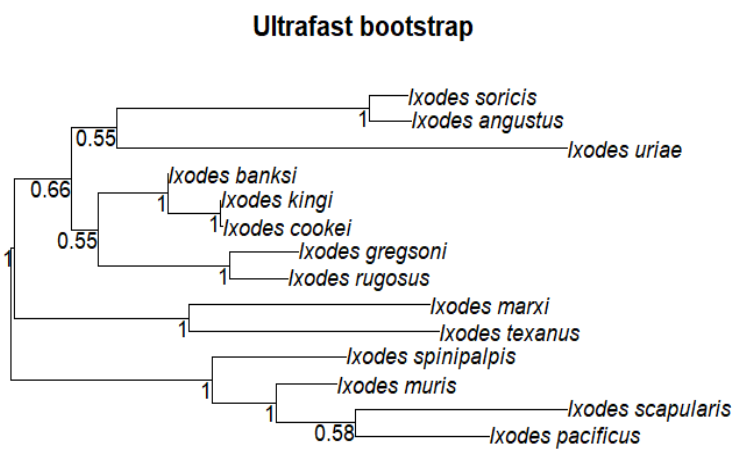



```
#compute likelihood data based on tree and evolutionary model
njTreeLikelihood <- pml(njTree, as.phyDat(alignmentBIN), k=4) #calculate
likelihood

#optimizetree
njTreeLikelihoodOptimized <- optim.pml(njTreeLikelihood, optNni=TRUE,
optBf=TRUE, optQ=TRUE, optGamma=TRUE)

#check if the optimization was significant
anova(njTreeLikelihood, njTreeLikelihoodOptimized) #it is significantly
different (p<0.001)

## Likelihood Ratio Test Table
##   Log lik. Df Df change Diff log lik. Pr(>|Chi|)
## 1  -3933.0 26
## 2  -3521.2 34         8        823.52  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Check AIC
AIC(njTreeLikelihood)

## [1] 7917.954

AIC(njTreeLikelihoodOptimized)

## [1] 7110.429

#lower AIC value is better, optimization was a success

#Check out the new tree
```

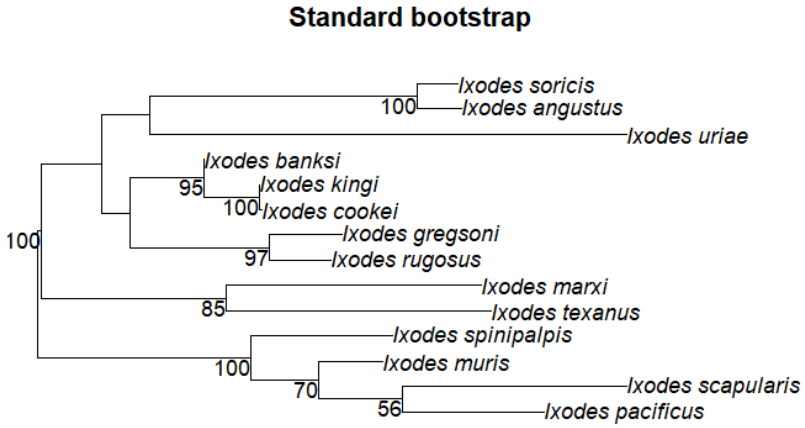
```
plot(njTreeLikelihoodOptimized, main="Ixodes UPGMA Neighbour Joining
Phylogeny Optimized")
```



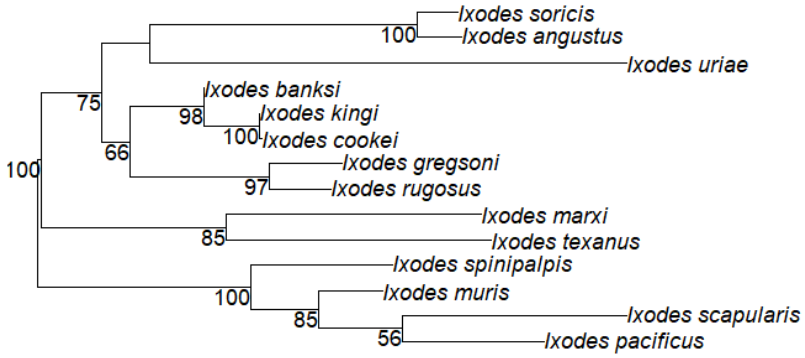

```
#Better than clustering!

###FIND BEST MODEL AND MODEL PHYLOGENY (NOT USED IN GRAPH)
#Other phylogeny methods, but failed to get correct data format for plotting

#try common models and determine best model
mt <- modelTest(as.phyDat(alignmentBIN), model=c("JC", "F81", "K80", "HKY",
"SYM", "GTR"), control = pml.control(trace = 0))
#save phylogeny from best model
best_mt <- pml_bb(mt, control = pml.control(trace = 0))

#calculate distances and plot
bs <- bootstrap.pml(best_mt, bs=100, optNni=TRUE, control = pml.control(trace
= 0))
plotBS(midpoint(best_mt$tree), p = .5, type="p", digits=2, main="Ultrafast
bootstrap")
```




```
plotBS(midpoint(best_mt$tree), bs, p = 50, type="p", main="Standard
bootstrap")
```




```
plotBS(midpoint(best_mt$tree), bs, p = 50, type="p", digits=0, method =
"TBE", main="Transfer bootstrap")
```



```r
#save standard bootstrap phylogeny
tree_stdbs <- plotBS(best_mt$tree, bs, type = "n")

###PLOT PHYLOGENY

#Simplify location data
dfcurated <- dfIxodesDataAvailable[ , c("species_name", "lat", "lon")]
dfcurated <- distinct(dfcurated, species_name, lat, lon) #remove repeats
row_names <- dfcurated$species_name #save species name
dfcurated$species_name <- NULL #remove species name
dfcurated <- as.matrix(dfcurated) #convert to matrix
row.names(dfcurated) <- row_names #add back species name as row names

#check that data is heretical and binary, and the species names match
is.ultrametric(phyloData)

## [1] TRUE

is.binary(phyloData)

## [1] TRUE

all.equal(phyloData$tip.label, unique(rownames(dfcurated)))

## [1] TRUE

#set colors
cols<-setNames(sample(viridis(n=Ntip(phyloData)), phyloData$tip.label)

#create geophylogeny map
tree_map <- phylo.to.map(tree = phyloData, coords = dfcurated, rotate = F,
type = "phylogram", regions = "Canada", plot = F)

#plot geophylogeny map
plot(tree_map, main="Geographical Distribution of Ixodes spp. in Canada",
xlim = c(-150, -50), ylim = c(40, 90), split = c(0.5, 0.5), fsize = 1, ftype
= "i", asp = 2, from.tip = F, map.bg = "azure", lwd = 0, pts = T, colors =
cols, cex.points = 1.5, delimit_map = T, direction="rightwards")

#plot geophylogeny map (zoomed in)
plot(tree_map, main="Geographical Distribution of Ixodes spp. in Canada",
xlim = c(-100, -50), ylim = c(40, 60), split = c(0.4, 0.5), fsize = 0.8,
ftype = "i", asp = 2, from.tip = F, map.bg = "azure", lwd = 0, pts = T,
colors = cols, cex.points = 1, delimit_map = T, direction="rightwards")
```

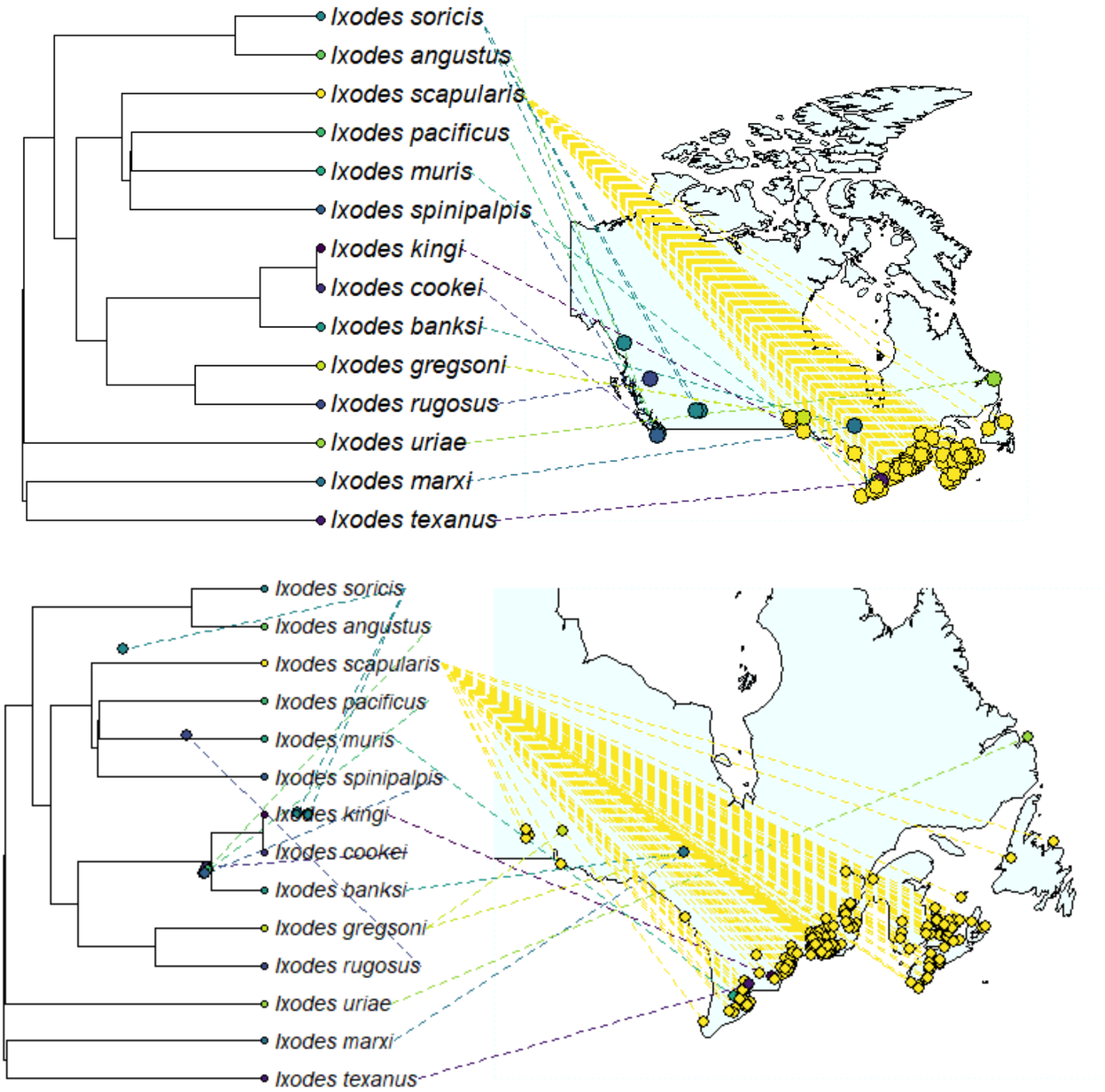


Figure 2. Geographical distribution of *Ixodes spp.* in Canada

Results and Discussion

In our analysis, we examined the genetic similarity between different species of *Ixodes* in Canada and further mapped these species to determine the proximity between related species of *Ixodes*. We can see from the final figure (fig. 2) that most species found in close proximity to our major species of concern (*Ixodes scapularis* as represented in yellow), such as *Ixodes kingi, Ixodes texanus, and Ixodes gregsoni,* are generally more distant in relation to *Ixodes scapularis* within the taxon. However, *Ixodes muris* being in the same clade as well as geological neighbors with *Ixodes scapularis* represents an area of concern. It may be possible that this species can also carry the *Borrelia burgdorferi* bacterium due to genetic similarities and may do so soon if they are frequently in the same area and share the same *Borrelia burgdorferi* infected food source as *Ixodes scapularis*. Similarly, our standard bootstrap phylogeny also supports this conclusion as it also suggests a close phylogenetic relationship between *Ixodes scapularis* and *Ixodes muris*. On the other hand, our optimized maximum likelihood phylogeny suggests instead that *Ixodes scapularis* is quite genetically different from other *Ixodes* species, thus proximity should not be cause for concern.

Overall, we determine that the greatest species of concern for gaining the ability to host *Borrelia burgdorferi* is *Ixodes muris*. However, it is necessary to confirm phylogenetic relationships between *Ixodes spp.* in Canada more rigorously by incorporating additional evolutionary marker genes. Additionally, one major limitation of this study was that we picked a random sequence to be representative of each species. An alternative could be to align the intra-species sequences and generate a consensus sequence for analysis instead. This could better represent species-level differences; however, we must also take care to not over-generalize the consensus sequences such that little to no inter-species variation can be detected. One other limitation is the

over-representation of *Ixodes scapularis* (over 300 entries) compared to other species, some of which are only represented by a single sequence (fig. 1). Interestingly, this has little to no impact on the strength of our conclusions due our randomized selection of a single sequence per species for alignment. However, when using the consensus sequence method described above, it may be pertinent to ensure a similar level of representation for each species analyzed. Lastly, all sequence data used in this analysis are sourced only from Canada and should not be generalized to other countries with their own distribution of *Ixodes spp.* such as Europe. Overall, this report represents an exploratory study of geophylogenic relationships between *Ixodes spp.* in Canada. In the future, experimental testing of the capacity of *Ixodes muris* to retain and carry *Borrelia burgdorferi* from consumed blood can be performed to definitively determine if *Borrelia burgdorferi* is able to colonize this host.

Reflection

I really enjoyed the visualization aspect of this course. I was able to learn, discover, and implement many colorful graphics and incorporate variety of way to represent data (eg. double linear plots, boxplots with scatter overlay, zooming on a section, plotting on a map, etc). I also enjoyed being able to pick our own subject topics. Although this proved frustrating at times as it was easy to lose direction or tunnel vision into an unrealistic project, some patience and research allowed me to explore many relevant and exciting topics such as marine speciation patterns with latitude away from equator, covid variant classifier, and Ixodes species responsible for Lyme disease spread in Canada. I think that I will be able to take my experience with the skill, scripts, and projects completed in this course to further explore machine learning and research in bioinformatics. Lastly, I was able to learn much from my peers. Generating ideas and receiving advice/feedback was very helpful for me, and something that I will continue in my future career.

Acknowledgements

I would like to acknowledge Sarah Adamowicz and our TA Jessica Castellanos Labarcena for the wonderful scripts and explanations provided in class. I would also like to thank my colleague Robin Zutshi for sharing his experience with geophylogeny graphing with me. I had faced an error with mismatching lengths between the phylogeny and geography data, and I received a suggestion to convert to matrix format and put species names as the row labels. I learned to examine the data format required closely and manipulate the data in strange ways to ensure format conformation. Finally, I would also like to thank stack overflow and the community for sharing ideas on how to solve particular problems (eg. remove x-axis ticks and labels).

Work Cited

Cayol, C., Jääskeläinen, A., Koskela, E., Kyröläinen, S., Mappes, T., Siukkola, A., & Kallio, E. R. (2018). Sympatric *Ixodes*-tick species: Pattern of distribution and pathogen transmission within wild rodent populations. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-35031-0

Edgar, R. C. (2004). Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340 Eisen, R. J., & Eisen, L. (2018). The blacklegged tick, *Ixodes scapularis* : An increasing public health concern. Trends in Parasitology, 34(4), 295–309. https://doi.org/10.1016/j.pt.2017.12.006

ES Wright (2016) "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." The R Journal, 8(1), 352-359.

Marvel, C. L., Alm, K. H., Bhattacharya, D., Rebman, A. W., Bakker, A., Morgan, O. P., Creighton, J. A., Kozero, E. A., Venkatesan, A., Nadkarni, P. A., & Aucott, J. N. (2022). A multimodal neuroimaging study of brain abnormalities and clinical correlates in post treatment lyme disease. PLOS ONE, 17(10). https://doi.org/10.1371/journal.pone.0271425

Revell L (2012). “phytools: An R package for phylogenetic comparative biology (and other things).” Methods in Ecology and Evolution, 3, 217-223. doi:10.1111/j.2041-210X.2011.00169.x.

Rizzoli, A., Silaghi, C., Obiegala, A., Rudolf, I., HubÃ¡lek, Z., FÃ¶ldvÃ¡ri, G., Plantard, O., Vayssier-Taussat, M., Bonnet, S., Å pitalskÃ¡, E., & KazimÃrovÃ¡, M. (2014). *Ixodes* ricinus and its transmitted pathogens in urban and Peri-Urban areas in Europe: New hazards and relevance for Public Health. Frontiers in Public Health, 2. https://doi.org/10.3389/fpubh.2014.00251

Schliep K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593

Slatculescu, A. M., Clow, K. M., McKay, R., Talbot, B., Logan, J. J., Thickstun, C. R., Jardine, C. M., Ogden, N. H., Knudby, A. J., & Kulkarni, M. A. (2020). Species distribution models for the eastern blacklegged tick, *Ixodes scapularis*, and the lyme disease pathogen, *Borrelia burgdorferi*, in Ontario, Canada. PLOS ONE, 15(9). https://doi.org/10.1371/journal.pone.023812