Evidencia 2

Juan Enrique Ayala Zapata-A01711235

2024-05-04

Parte 1: Video.

Link: https://youtu.be/H10vpvGTBvs

Parte 2: Codigo:

Opcion a analizar: cercanía del SARS-CoV-2 con otros coronavirus humanos.

Importar librerias y asignar el espacio de trabajo.

```
library(ade4)
library(ape)
library(adegenet)
```

```
##
##    /// adegenet 2.1.10 is loaded ////////////
##
##    > overview: '?adegenet'
##    > tutorials/doc/questions: 'adegenetWeb()'
##    > bug reports/feature requests: adegenetIssues()
```

```
library(Biostrings)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following object is masked from 'package:ade4':
##
##     score
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
```

```
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:ape':
##
##     complement

## The following object is masked from 'package:base':
##
##     strsplit

library(ggplot2)
library(stringr)
```

Longitud de las secuencias:

```
variantes <- c("Civet.fasta", "HCoV-229E.fna", "HCoV-HKU1.fna", "HCoV-
NL63.fna", "HCoV-OC43.fna", "MERS-CoV.fna", "RaTG13.fasta",
"RpYN06.fasta", "SARS-CoV.fna", "SARS-CoV-2.fna")
nombre_variantes <- c("Civet SARS-CoV", "Human Coronavirus 229E", "Human
Coronavirus HKU1", "Human Coronavirus NL63", "Human Coronavirus OC43",
"Middle East Respiratory Syndrome", "Bat Coronavirus RaTG13",
```

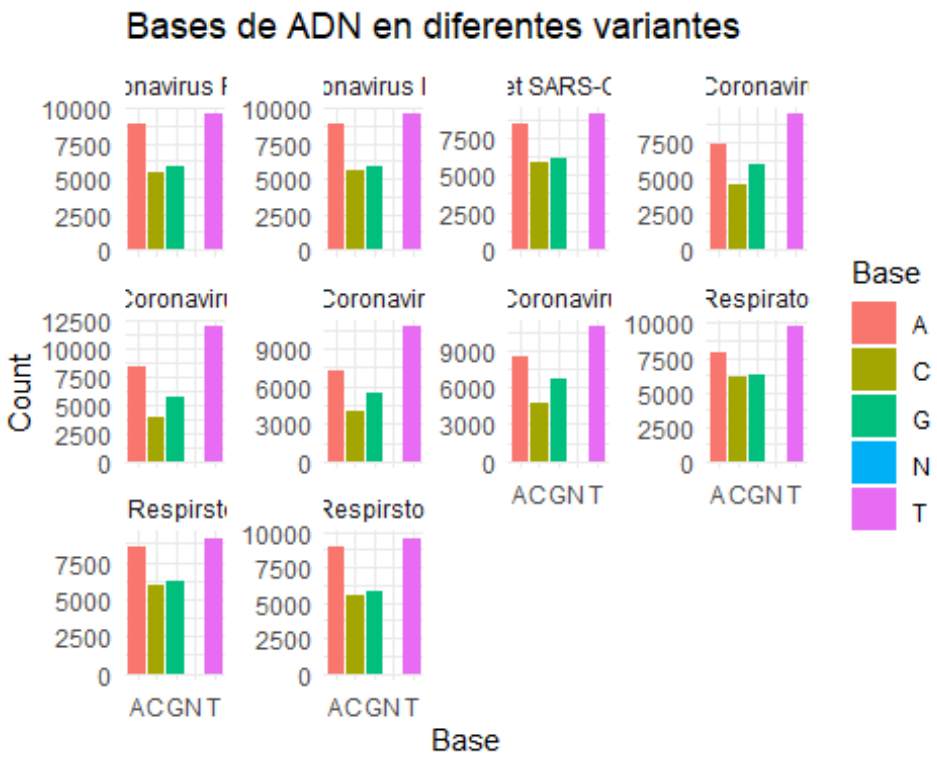
```
"Betacoronavirus RpYN06", "Severe Acute Respirstory Syndrome", "Severe
Acute Respirstory Syndrome 2")

contar_bases <- function(ADN) {
  a_count <- str_count(ADN, pattern = "A")
  t_count <- str_count(ADN, pattern = "T")
  g_count <- str_count(ADN, pattern = "G")
  c_count <- str_count(ADN, pattern = "C")
  n_count <- str_count(ADN, pattern = "N")
  data.frame(Base = c("A", "T", "G", "C", "N"), Count = c(a_count,
t_count, g_count, c_count, n_count))
}

datos <- list()
nombres <- list()

for (i in 1:length(variantes)) {
  ADN_set <- readDNAStringSet(variantes[i])
  adn <- toString(ADN_set)
  base_counts <- contar_bases(adn)

  datos[[i]] <- base_counts
  nombres[[i]] <- nombre_variantes[i]
}

datos_combinados <- do.call(rbind, datos)

datos_combinados$Variante <- rep(nombre_variantes, each = 5)

dna <- fasta2DNAbin(file="usflu.fasta")

##
##  Converting FASTA alignment into a DNAbin object...
##
##
##  Finding the size of a single genome...
##
##
##  genome size is: 29,540 nucleotides
##
## ( 423  lines per genome )
##
##  Importing sequences...
## ..........

## Warning in split.default(txt, rep(1:nb.ind, each = LINES.PER.IND)):
largo de
## datos no es múltiplo de la variable de separación
```

```
##
##  Forming final object...

## Warning in matrix(res, nrow = length(IND.LAB), byrow = TRUE): la
longitud de
## los datos [293923] no es un submúltiplo o múltiplo del número de filas
[10] en
## la matriz

##
## ...done.

dna

## 10 DNA sequences in binary format stored in a matrix.
##
## All sequences of same length: 29393
##
## Labels:
## CIVET
## 229E
## HKU1
## NL63
## OC43
## MERS
## ...
##
## Base composition:
##     a     c     g     t
## 0.282 0.175 0.204 0.339
## (Total: 293.93 kb)
```

Grafica de comparación de numero de ADN.

```
contar_bases <- function(ADN) {
  a_count <- str_count(ADN, pattern = "A")
  t_count <- str_count(ADN, pattern = "T")
  g_count <- str_count(ADN, pattern = "G")
  c_count <- str_count(ADN, pattern = "C")
  n_count <- str_count(ADN, pattern = "N")
  data.frame(Base = c("A", "T", "G", "C", "N"), Count = c(a_count,
t_count, g_count, c_count, n_count))
}

datos <- list()
nombres <- list()

for (i in 1:length(variantes)) {
  ADN_set <- readDNAStringSet(variantes[i])
  adn <- toString(ADN_set)
  base_counts <- contar_bases(adn)
```

```r
  datos[[i]] <- base_counts
  nombres[[i]] <- nombre_variantes[i]
}

datos_combinados <- do.call(rbind, datos)

datos_combinados$Variante <- rep(nombre_variantes, each = 5)

ggplot(datos_combinados, aes(x = Base, y = Count, fill = Base)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Variante, scales = "free_y") +
  labs(title = "Bases de ADN en diferentes variantes",
       x = "Base", y = "Count") +
  theme_minimal()
```

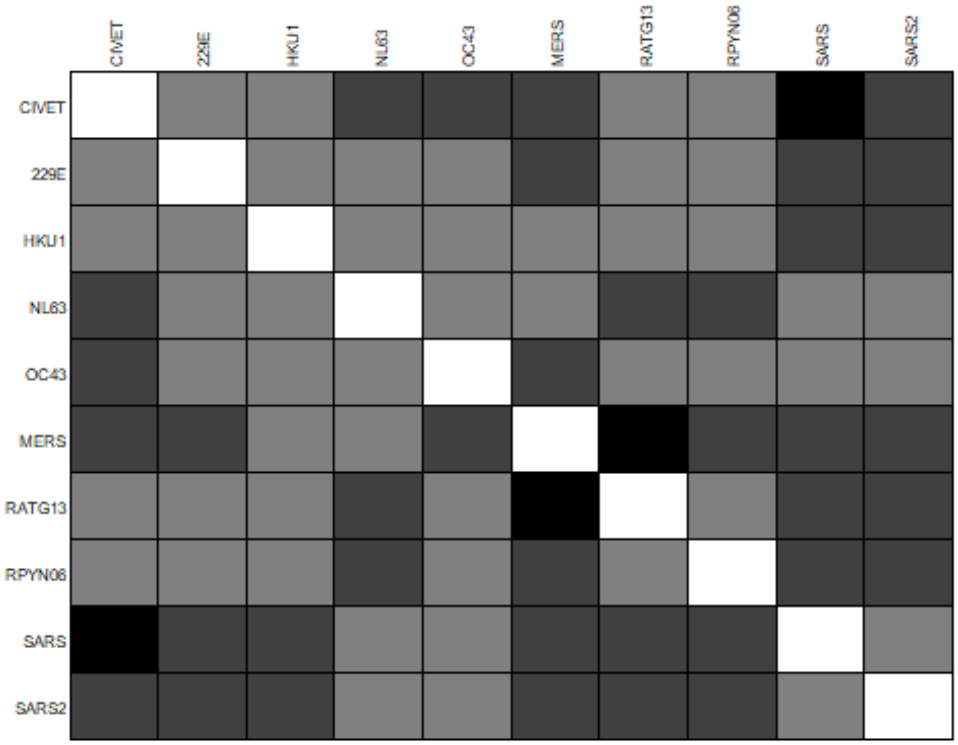


Analisis jerárquico de las secuencias.

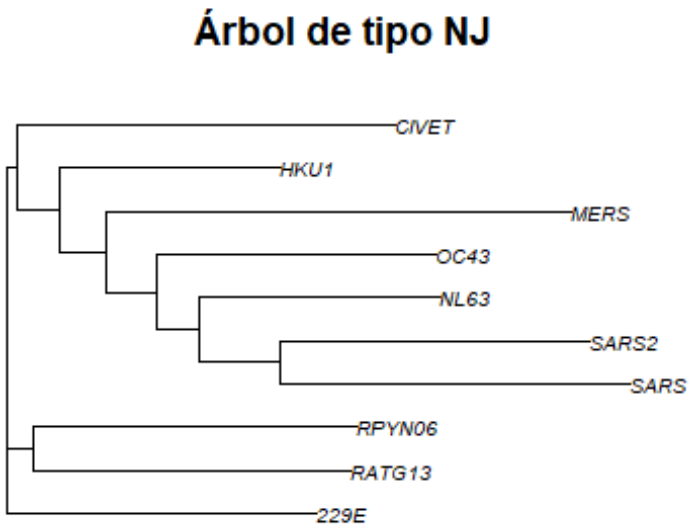
Elección de método JC69

El modelo JC69 asume que todas las sustituciones tienen la misma probabilidad (cambiar una base por otra). Esta probabilidad es la misma para todos los sitios a lo largo de la secuencia de ADN. Esta última observación está relacionada con la hipótesis de que todsa las sustituciones varían de acuerdo el sitio siguiendo la distribución GAMMA, cuyo parámetro debe de ser dado por el usuario.

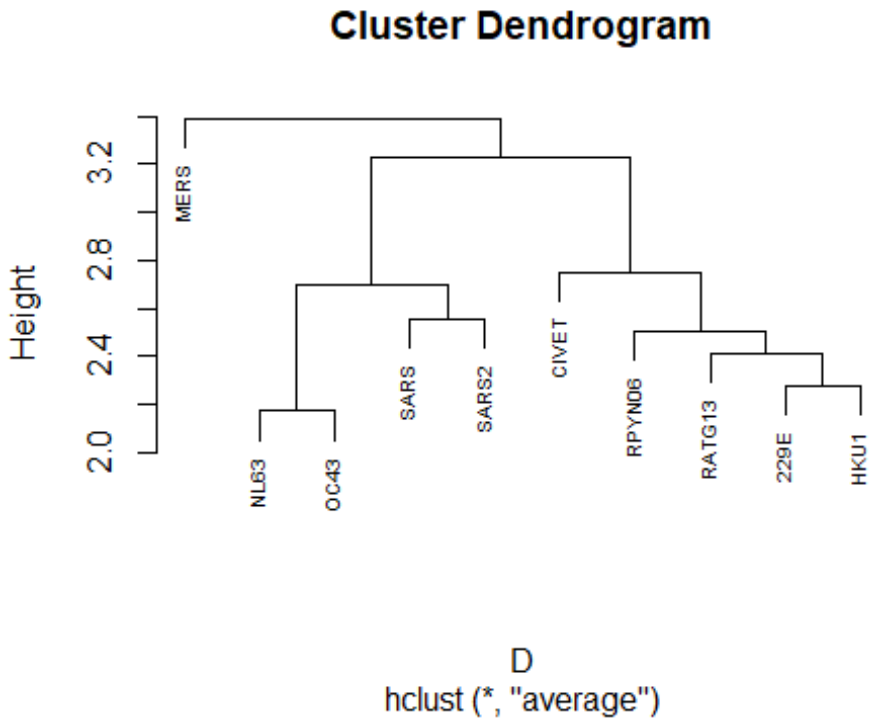
```
annot <- read.csv("usflu.annot.csv", header=TRUE, row.names=1)

D <- dist.dna(dna, model = "JC69")
length(D)

## [1] 45

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

MatrizDG <- as.data.frame(as.matrix(D))
table.paint(MatrizDG, cleg = 0, clabel.row = 0.5, clabel.col = 0.5)
```



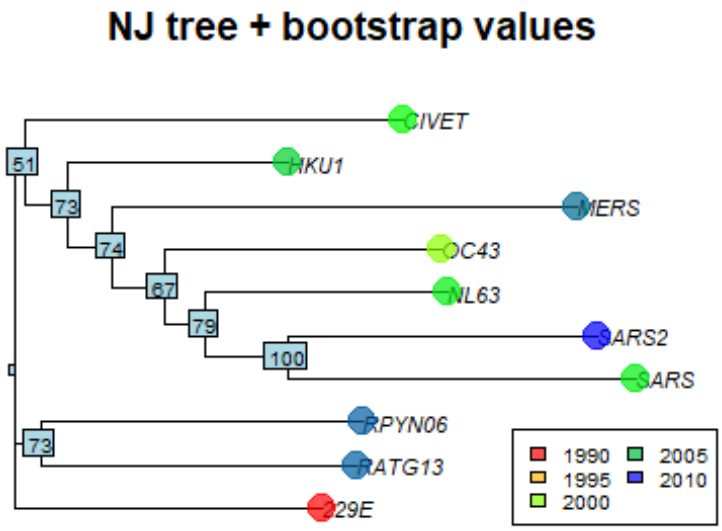

```
tre <- njs(D)
plot(tre, cex = 0.6)
title("Árbol de tipo NJ")
```




```r
h_cluster <- hclust(D, method = "average", members = NULL)
plot(h_cluster, cex = 0.6)
```



```
myBoots <- boot.phylo(tre, dna, function(e) root(njs((dist.dna(e, model =
"JC69"))), 1))

## Running bootstraps:       100 / 100
## Calculating bootstrap values... done.

myPal <- colorRampPalette(c("red", "yellow", "green", "blue"))
plot(tre, edge.width = 1, cex = 0.7)
title("NJ tree + bootstrap values")
tiplabels(frame = "none", pch = 20, col = transp(num2col(annot$year,
col.pal = myPal), 0.7), cex = 3, fg = "transparent")
temp <- pretty(1993:2008, 5)
legend("bottomright", fill = transp(num2col(temp, col.pal = myPal), 0.7),
leg = temp, ncol = 2, cex = 0.6)
nodelabels(myBoots, cex = 0.6)
```



Matriz distancia:

En la matriz distancia podemos observar como todos los cuadros de esta, están coloreados dependiendo de la similitud entre cada secuencia de ADN de cada variante del virus CoV a analizar en este proyecyo. Que tan “negro” esté coloreado el cuadro, significa que tan cercano es el virus respecto al otro.

NJ Tree

Este árbolo filogenético nos presenta las relaciones evolutivas entre los diferentes virus de la familia Coronaviridae escogidos a ser analizados en este proyecto, la longitud de cada rama representa la cantidad de tiempo que ha pasado desde que dos organismos divergieron de un ansestro común. Así mismo, las secuencias de ADN que se encuentran en la misma rama, están más relacionados entre sí que con los organismos que se encuentran en otras ramas.

Cluster Dendogram

Nos muestra la relación entre la altura de los diferentes grupos de la familia Coronaviridae. Cuanto más corta sea la rama, más similiar es el grupo de variantes que representa. Cuanto más larga, más diferente es el grupo representado. El punto de fusión de cada rama representa la altura en que los dos grupos se fusionan en un solo grupo. Cuanto más alto sea este punto, más diferentes son los grupos.

NJ Tree + Bootstrap Values

Este árbolo filogenético nos presenta las relaciones evolutivas entre los diferentes virus de la familia Coronaviridae escogidos a ser analizados en este proyecto, la longitud de cada rama representa la cantidad de tiempo que ha pasado desde que dos organismos divergieron de un ansestro común. Así mismo, las secuencias de ADN que se encuentran en la misma rama, están más relacionados entre sí que con los organismos que se encuentran en otras ramas. Los valores de bootstrap que se muestran, representan la confianza que se tiene en la posición de esos nodos. Estos valores, en este árbol son elevados, lo que indica que hay una alta confianza en las relaciones evolutivas presentadas en este árbol.

Referencias

Coronavirus disease (COVID-19) pandemic. (2024, 1 mayo). https://www.who.int/europe/emergencies/situations/covid-19

Islam, A., Ferdous, J., Sayeed, M. A., Islam, S., Rahman, M. L., Abedin, J., Saha, O., Hassan, M. M., & Shirin, T. (2021). Spatial epidemiology and genetic diversity of SARS-CoV-2 and related coronaviruses in domestic and wild animals. PloS One, 16(12), e0260635. https://doi.org/10.1371/journal.pone.0260635

Ge, X., Wang, N., Zhang, W., Hu, B., Li, B., Zhang, Y. Z., Zhou, J., Luo, C. M., Yang, X., Li, W., Wang, B., Zhang, Y., Li, Z. X., & Shi, Z. L. (2016). Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. Virologica Sinica/Virologica Sinica, 31(1), 31-40. https://doi.org/10.1007/s12250-016-3713-9

Pavan, M., Bassani, D., Sturlese, M., & Moro, S. (2022). Bat coronaviruses related to SARS-CoV-2: what about their 3CL proteases (MPro)? Journal Of Enzyme Inhibition

And Medicinal Chemistry, 37(1), 1077-1082.
https://doi.org/10.1080/14756366.2022.2062336

SARS-CoV-2, SARS-CoV, and MERS-COV: A comparative overview. (2020). PubMed.
https://pubmed.ncbi.nlm.nih.gov/32275259/