Tecnológico de Monterrey

INSTITUTO TECNOLÓGICO DE

ESTUDIOS SUPERIORES DE MONTERREY

Análisis de biología computacional (Gpo 856)

“Evidencia 2| Proyecto integrador”

Situación Problema

29 de Abril del 2022

Adrian Pineda Sánchez **A00834710**

Arturo Ramos **A01636133**

Código

1- Diseña un análisis filogenético del virus SARS-CoV-2, en donde incluyas de 8 a 10 genomas virales.

```r
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("Biostrings")

if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DECIPHER")

library("Biostrings")
library("DECIPHER")
library("seqinr")

#Lectura
omiFrancia <- read.fasta("Omicron_Francia.fasta")
delIndia <- read.fasta("Delta_India.fasta")
omiIndia <- read.fasta("Omicron_India.fasta")
alphaUSA <- read.fasta("Alpha_USA.fasta")
omiUSA <- read.fasta("Omicron_USA.fasta")
omiOCEA <- read.fasta("Omicron_OCEA.fasta")
alphaSDA <- read.fasta("Alpha_SDA.fasta")
delSDA <- read.fasta("Delta_SDA.fasta")

#Longitud del genoma
omiFranciaLength <- length(omiFrancia[[1]])
delIndiaLength <- length(delIndia[[1]])
omiIndiaLength <- length(omiIndia[[1]])
alphaUSALength <- length(alphaUSA[[1]])
omiUSALength <- length(omiUSA[[1]])
omiOCEALength <- length(omiOCEA[[1]])
alphaSDALength <- length(alphaSDA[[1]])
delSDALength <- length(delSDA[[1]])

genomasLongitudes <- c(omiFranciaLength, delIndiaLength, omiIndiaLength, alphaUSALength, omiUSALength, omiOCEALength, alphaSDALength, delSDALength)

#Contenido de GC
omiFranciaGC <- GC(omiFrancia[[1]])
delIndiaGC <- GC(delIndia[[1]])
omiIndiaGC <- GC(omiIndia[[1]])
alphaUSAGC <- GC(alphaUSA[[1]])
```

```r
omiUSAGC <- GC(omiUSA[[1]])
omiOCEAGC <- GC(omiOCEA[[1]])
alphaSDAGC <- GC(alphaSDA[[1]])
delSDAGC <- GC(delSDA[[1]])

genomasGC <-c(omiFranciaGC, delIndiaGC, omiIndiaGC, alphaUSAGC, omiUSAGC, omiOCEAGC, alphaSDAGC, delSDAGC)

aPorcentajeLista = c()
cPorcentajeLista = c()
tPorcentajeLista = c()
gPorcentajeLista = c()
nombres = c()

composicion <- function(genomaVirus, nombre){

  genoma = readDNAStringSet(genomaVirus)
  genoma

  ancho <- width(genoma)
  frecuencia <- alphabetFrequency(genoma, baseOnly = TRUE)

  aFrec <- frecuencia[1]
  cFrec <- frecuencia[2]
  gFrec <- frecuencia[3]
  tFrec <- frecuencia[4]

  cat("A: ", aFrec,
      "\nC: ", cFrec,
      "\nG: ", gFrec,
      "\nT: ", tFrec
  )

  aPorcentaje <- (aFrec/ancho) * 100
  cPorcentaje <- (cFrec/ancho) * 100
  gPorcentaje <- (gFrec/ancho) * 100
  tPorcentaje <- (tFrec/ancho) * 100

  cat("\n% de A: ", aPorcentaje,
      "\n% de C: ", cPorcentaje,
      "\n% de G: ", gPorcentaje,
```

```
      "\n% de T: ", tPorcentaje)


  aPorcentajeLista <<- c(aPorcentajeLista, aPorcentaje)
  cPorcentajeLista <<- c(cPorcentajeLista, cPorcentaje)
  tPorcentajeLista <<- c(tPorcentajeLista, tPorcentaje)
  gPorcentajeLista <<- c(gPorcentajeLista, gPorcentaje)
  nombres <<- c(nombres, nombre)

}

composicion("Omicron_Francia.fasta", "Omicron-Francia")
composicion("Delta_India.fasta", "Delta-India")
composicion("Omicron_India.fasta", "Omicron-India")
composicion("Alpha_USA.fasta", "Alpha-USA")
composicion("Omicron_USA.fasta", "Omicron-USA")
composicion("Omicron_OCEA.fasta", "Omicron-Oceanía")
composicion("Alpha_SDA.fasta", "Alpha-Sudáfrica")
composicion("Delta_SDA.fasta", "Delta-Sudáfrica")



par(mar=c(7,4,4,4))
barplot(aPorcentajeLista,
        horiz = FALSE,
        names.arg = nombres,
        axes = TRUE,
        xlab="",
        ylab="Porcentaje",
        las = 2,
        cex.names = 0.8,
        cex.axis = 0.8,
        ylim = c(29.8, 29.95),
        xpd = FALSE,
        col="#ff4040",
        main = "Porcentaje de Adenina",
        border="black")

par(mar=c(7,4,4,4))
barplot(cPorcentajeLista,
        horiz = FALSE,
        names.arg = nombres,
```

```
        axes = TRUE,
        xlab="",
        ylab="Porcentaje",
        las = 2,
        cex.names = 0.8,
        cex.axis = 0.8,
        ylim = c(18.27, 18.38),
        xpd = FALSE,
        col="#9cff45",
        main = "Porcentaje de Citocina",
        border="black")

par(mar=c(7,4,4,4))
barplot(tPorcentajeLista,
        horiz = FALSE,
        names.arg = nombres,
        axes = TRUE,
        xlab="",
        ylab="Porcentaje",
        las = 2,
        cex.names = 0.8,
        cex.axis = 0.8,
        ylim = c(32.07, 32.23),
        xpd = FALSE,
        col="#ffd045",
        main = "Porcentaje de Timina",
        border="black")

par(mar=c(7,4,4,4))
barplot(gPorcentajeLista,
        horiz = FALSE,
        names.arg = nombres,
        axes = TRUE,
        xlab="",
        ylab="Porcentaje",
        las = 2,
        cex.names = 0.8,
        cex.axis = 0.8,
        ylim = c(19.58, 19.65),
        xpd = FALSE,
        col="#3eb8f0",
        main = "Porcentaje de Guanina",
```

```
par(mar=c(7,4,4,4))
barplot(genomasGC,
        horiz = FALSE,
        names.arg = nombres,
        axes = TRUE,
        xlab="",
        ylab="GC",
        las = 2,
        cex.names = 0.8,
        cex.axis = 0.8,
        ylim = c(0.378, 0.381),
        xpd = FALSE,
        col="#0096ff",
        main = "Contenido GC",
        border="black")
```

```
virus_seq_align <- AlignSeqs(virus_seq_not_align)
BrowseSeqs(virus_seq_align)

writeXStringSet(virus_seq_align, file="coronavirus_sec_align.fasta")

virus_aligned <- read.alignment("coronavirus_sec_align.fasta", format="fasta")

matriz_distancia <- dist.alignment(virus_aligned, matrix = "similarity")
matriz_distancia_frame <- as.data.frame(as.matrix(matriz_distancia))
tabla_grises <- table.paint(matriz_distancia_frame, cleg=0, clabel.row=.5, clabel.col=.5) + scale_color_viridis()

virus_tree <- nj(matriz_distancia)

virus_tree <- ladderize(virus_tree)
plot(virus_tree)
```

2- Calcula la longitud y el contenido GC de las secuencias que incluyes en tu análisis. Crea una gráfica que te permita visualizar y comparar la longitud de los genomas analizados y otra que te permita ver el contenido de GC. Muestra el código empleado para obtenerlo e incluye las gráficas que obtuviste.

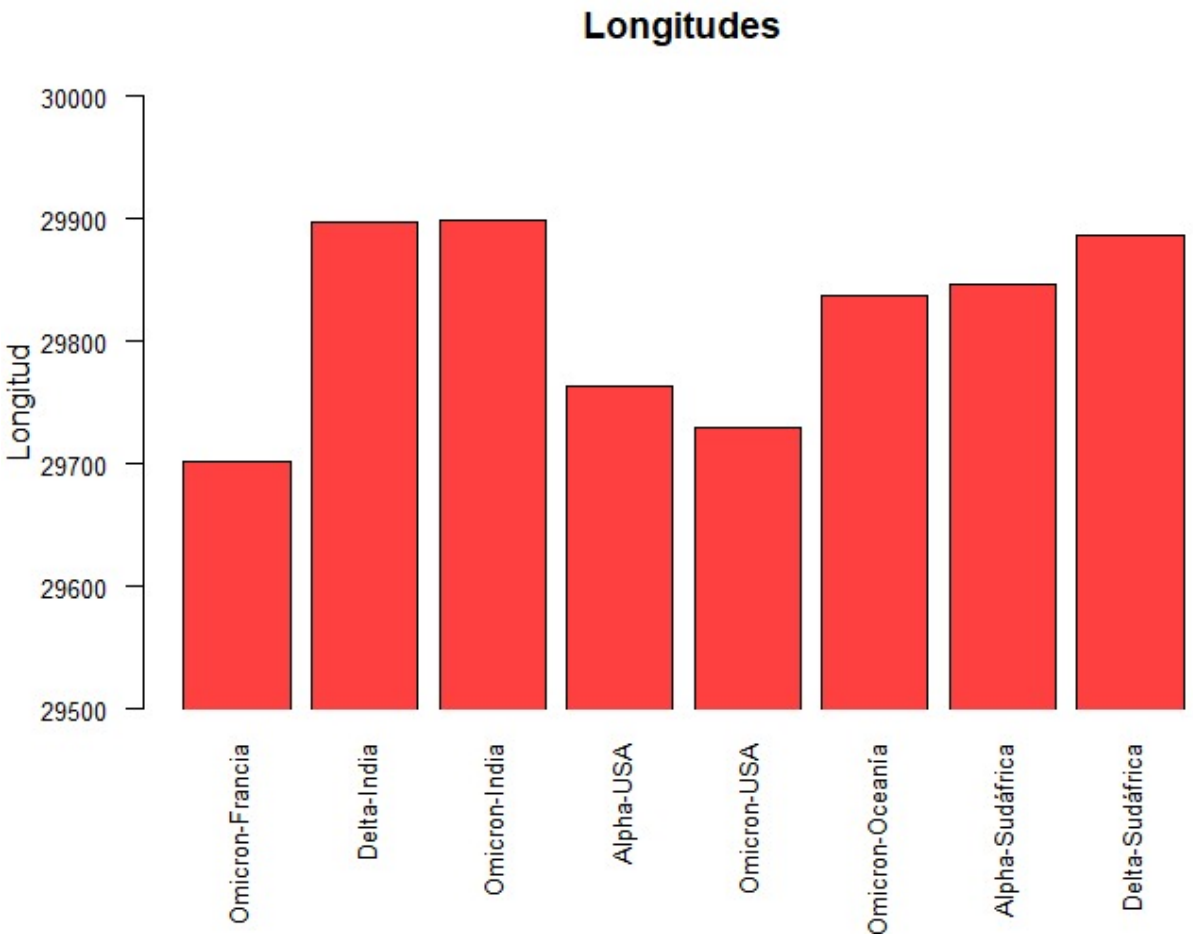

Figura 1.1 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica la longitud de las secuencias de las variantes del virus Sars Cov-2

Podemos observar a través de la **figura 1.1**, que la longitud mayor, aún y cuando es casi imperceptible entre las variantes Delta y Ómicron de la India, Ómicron parece encabezar la

lista sobre el resto de variantes, y en puestos posteriores podemos encontrar a las variantes Alpha y Delta de Sudáfrica, posterior a Ómicron de Oceanía, seguido de Alpha y Ómicron de los Estados Unidos y finalmente terminando con Ómicron de Francia, algo destacable, es que las variantes procedentes del mismo origen territorial parecen compartir una similitud entre la relación perceptible de la longitud de su genoma, debido a que las variantes ordenadas de mayor a menor destacan por estar distribuidas en pares con su país de origen y no de forma aleatoria.

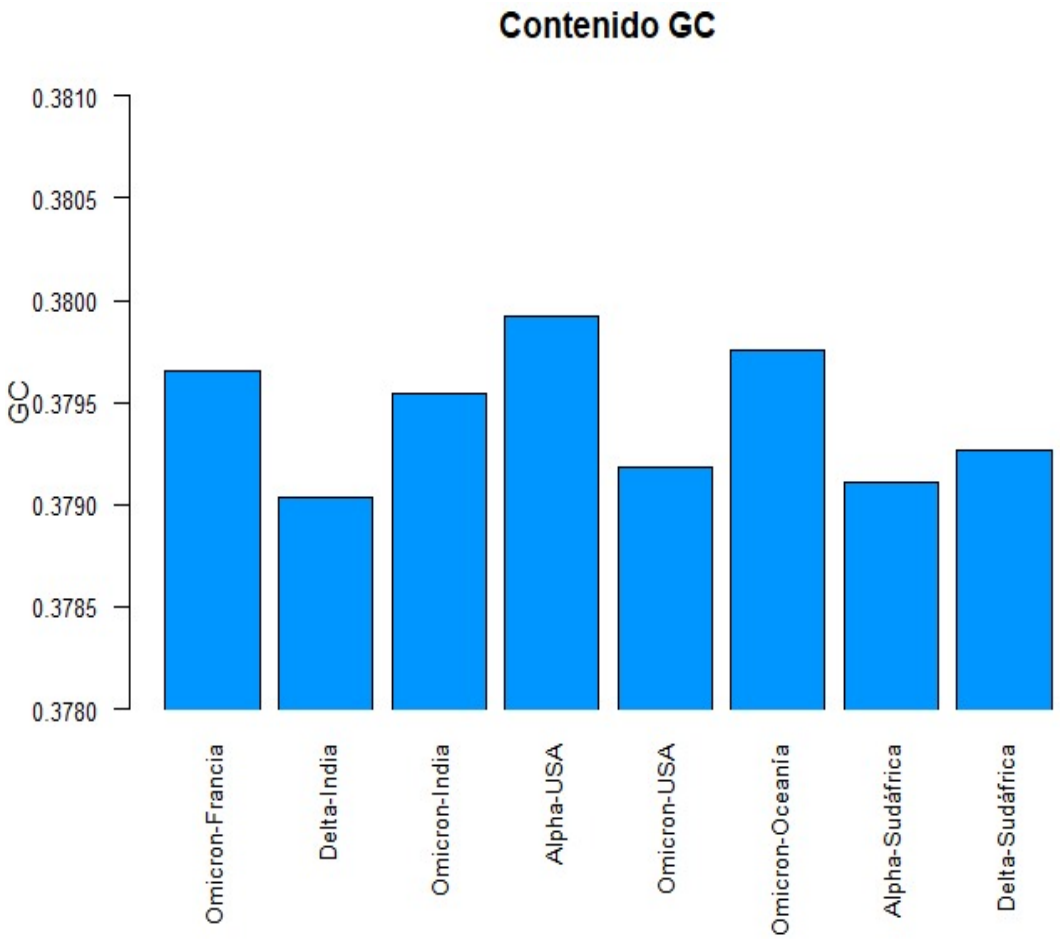

Figura 1.2 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica la cantidad porcentual (siendo 1 el 100 por ciento) del contenido de GC de las secuencias variantes del virus Sars Cov-2

A través de la **figura 1.2** podemos observar la distribución porcentual de GC a la cual se encuentran las distintas variantes estudiadas, en las que podemos destacar el encabezamiento de la variante Alpha de Estados Unidos por encima de las otra, seguida de Ómicron procedente de Oceanía y posterior a Ómicron de Francia, seguido a continuación por Delta de

Sudáfrica, Ómicron de Estados Unidos, Alpha de Sudáfrica y finalmente con Delta de India, aun y cuando podríamos decir que se tiene un cambio significativamente notable por lo mostrado en las gráficas, hay que considerar la escala manejada a un nivel demasiado preciso, ya que los intervalos están a milésimas porcentuales de distancia unos de otros, variando en un rango bastante ajustado, por lo que podemos destacar que dichas variantes presentan un punto en común en dicha concentración de GC.

3- Crea una gráfica en donde se observe el número de bases (A, G, C y T) de ADN que componen a los genomas virales utilizados en tu análisis. Muestra el código empleado para obtenerlo y la imagen de las gráficas obtenidas.

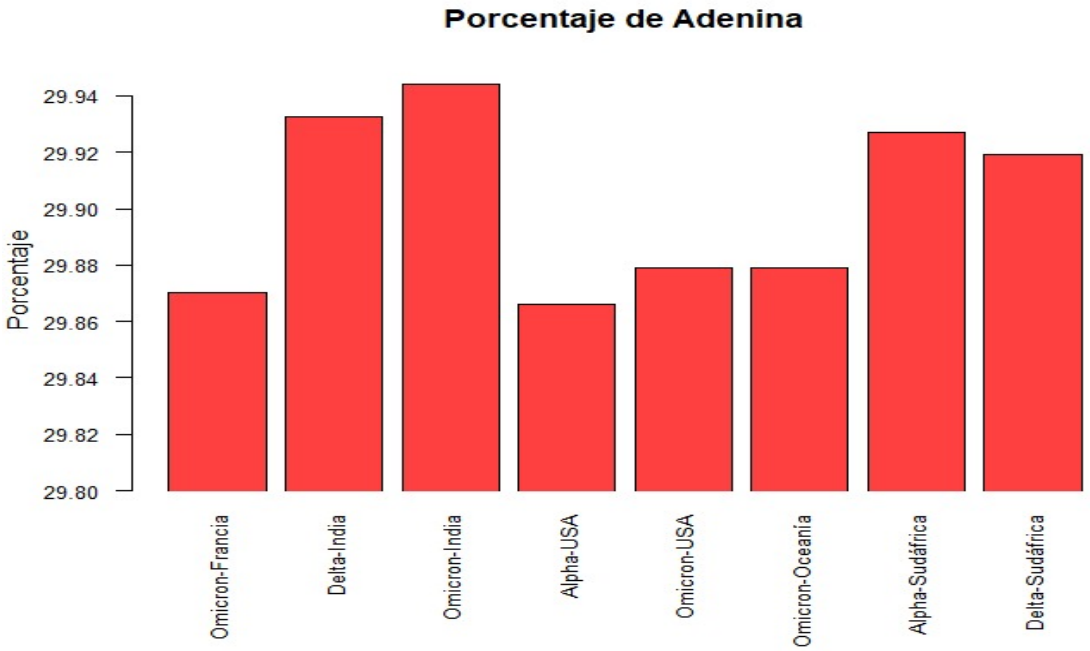

Figura 1.3 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica el contenido de Adenina de las secuencias variantes del virus Sars Cov-2

Podemos observar en la **figura 1.3** la distribución de la base Adenina en las distintas variantes estudiadas, podemos observar cómo Ómicron-India encabeza la lista posicionándose primera en dicha concentración, seguida de igual medida de Delta-India, posteriormente Alpha y Delta de Sudáfrica, seguida de Ómicron de Oceanía y posterior a ella Ómicron de Estados Unidos, después Ómicron de Francia y terminando con Alpha de Estados Unidos, podemos observar cómo al igual que en la Figura 1.2 y como en todas las gráficas posteriores de distribución de bases, el rango al cual todas las variantes varían es

muy limitado, siendo necesario utilizar una mayor cantidad de cifras significativas en cuestión decimal para poder apreciarse a un punto de vista gráfico.

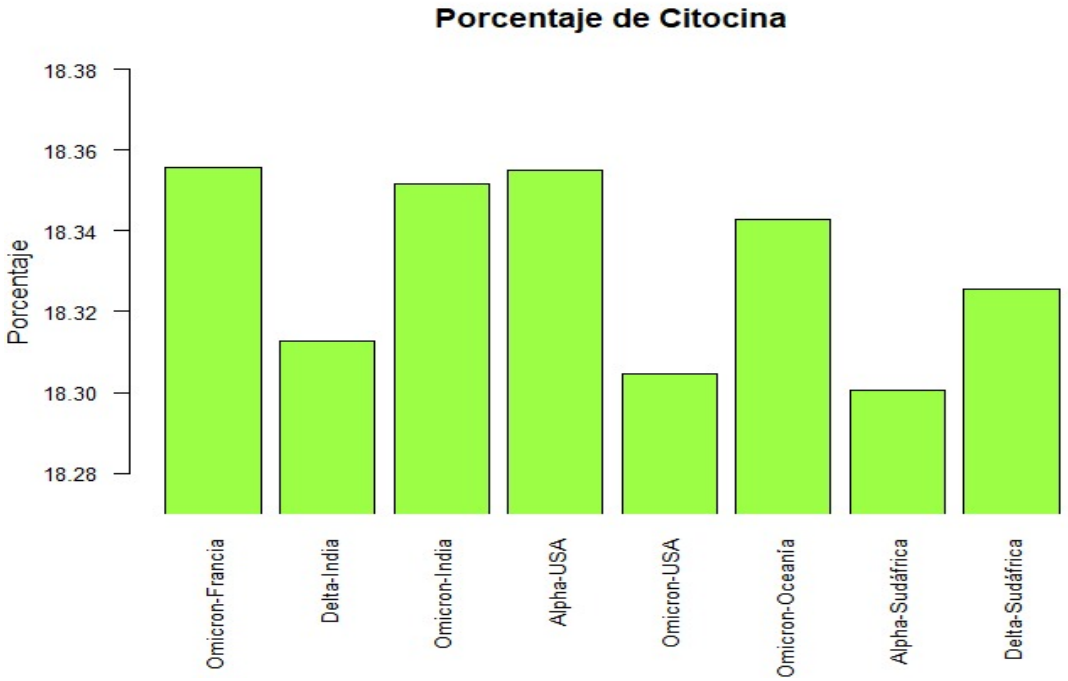

Figura 1.4 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica el contenido de Citocina de las secuencias variantes del virus Sars Cov-2

La **figura 1.4** nos describe la distribución porcentual de la base citocina en las distintas variantes, en la cual podemos destacar el encabezamiento de la variante Ómicron de Francia en esta lista, seguido de Alpha de Estados Unidos, así como de Ómicron de India, posterior a ellos Ómicron de Oceanía, Delta de Sudáfrica, Delta de India, Ómicron de Oceanía y Alpha de Sudáfrica, recalcando lo dicho anteriormente en las interpretaciones anteriores en dicho informe en cuestión a la distribución de las bases de cada variante, el rango al cual varían en cuestión porcentual es decimal, por lo que todas se encuentran en un entorno demasiado cercano en cuanto a la distribución de sus bases.

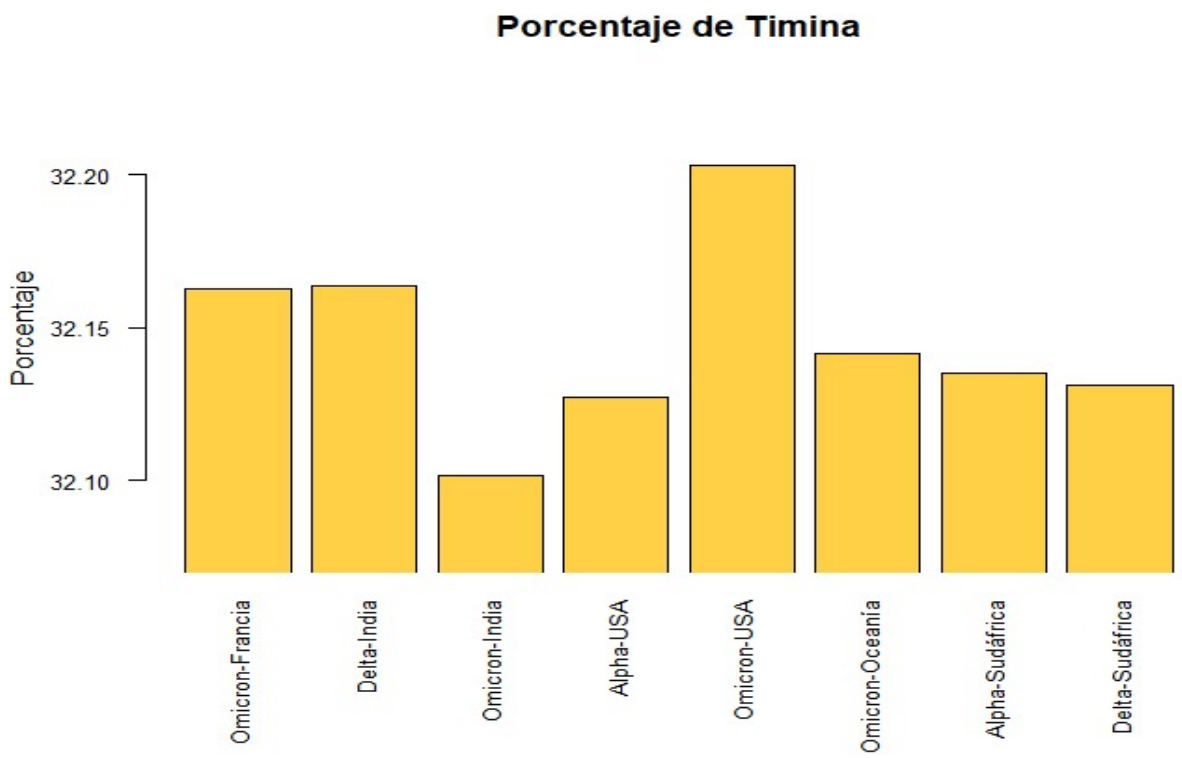

Figura 1.5 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica el contenido de Timina de las secuencias variantes del virus Sars Cov-2

En la **figura 1.5** podemos observar la distribución de la base Timina en las distintas variantes de objeto de estudio, encabezando la lista se encuentra la variante Ómicron de Estados Unidos, seguido de Delta de India y Ómicron de Francia, posterior a ellas Ómicron de Oceanía, Alpha y Delta de Sudáfrica y finalmente Alpha de Estados Unidos y Ómicron de India, recalcando lo dicho anteriormente, la variabilidad a un nivel significativo es casi imperceptible a menos que nos enfoquemos a varias cifras significativas decimales, por lo que reforzamos la idea que todas las variantes comparten una distribución muy similar entre sus bases.

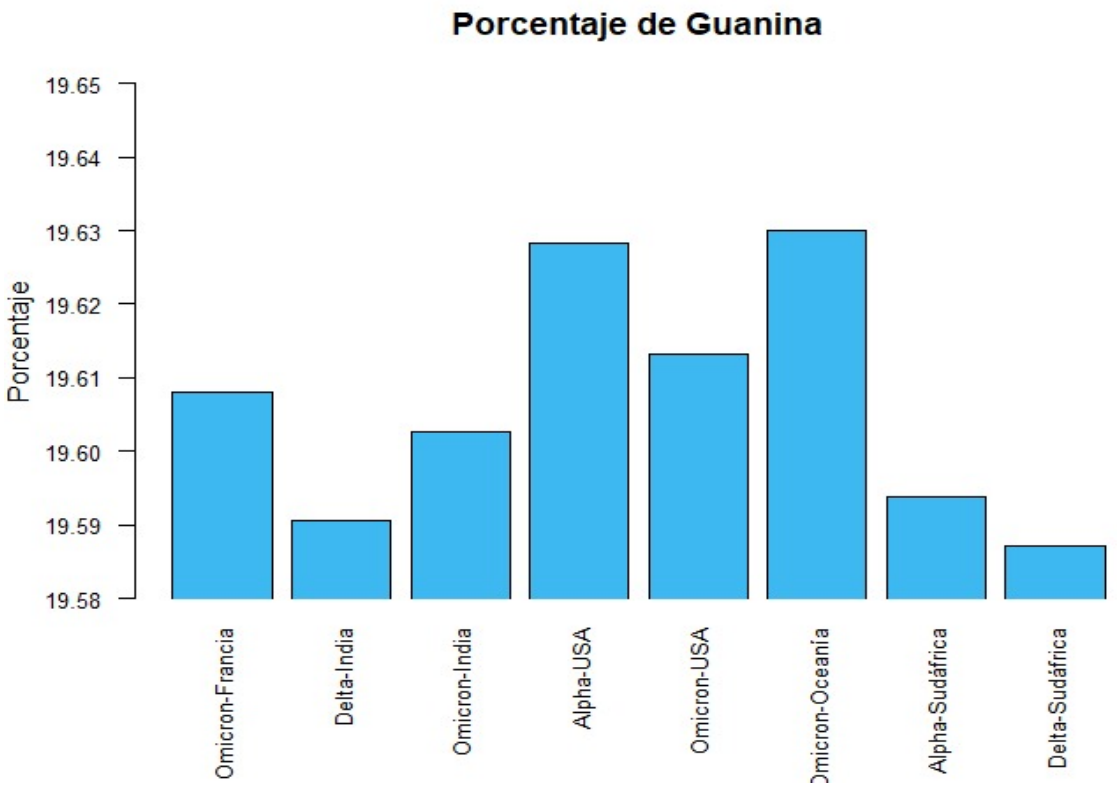

Figura 1.6 Gráfica obtenida a partir de la IDE Rscript donde se plasma de forma gráfica el contenido de Guanina de las secuencias variantes del virus Sars Cov-2

En la **figura 1.6** podemos vislumbrar nuestra última base a describir que es la Guanina, en cuestión de su distribución a través de las distintas variantes estudiadas, y reforzando la idea que se planteaba en las descripciones de las **figuras 1.3, 1.4 y 1.5**, podemos observar que en todas sus bases las variantes parecen guardar una correlación entre la distribución porcentual de cada una de ellas, debido a que mantienen un margen de variabilidad demasiado estrecho, teniendo que ser demasiado específicos, observando varios decimales o cifras significativas después del punto para vislumbrar un cambio a nivel grafico por ejemplo. En este sistema podemos destacar a Ómicron de Oceanía encabezando la lista, seguido de Alpha y Ómicron de Estados Unidos, Ómicron de Francia, Ómicron de India, Alpha y Delta de Sudáfrica y finalmente Delta de India.

4- Obtén las secuencias de los genomas de los virus elegidos, según la investigación que hayas decidido realizar, con la función read.GenBank. Muestra el código empleado para obtenerlo

```
> virus_sequences
8 DNA sequences in binary format stored in a list

Mean sequence length: 29820.12
   Shortest sequence: 29702
    Longest sequence: 29899

Labels:
ON287427
OK189630
MZ336026
ON134749
ON188698
OM737996
...

Base composition:
    a     c     g     t
0.299 0.183 0.196 0.322
(Total: 238.56 kb)
```

Figura 1.6 Imagen obtenida a partir de la IDE Rscript donde se plasman las secuencias de los genomas de las variantes del virus Sars Cov-2

La **figura 1.6** muestra el resultado de desplegar la variable virus_sequences en consola. Esta variable contiene el resultado de ejecutar la función read.GenBank de la librería ape usando nuestras variantes como argumentos. Podemos observar el número de secuencias almacenadas en la lista, además de la secuencia más larga, la más corta, y la longitud promedio de todas las secuencias. También nos despliega los códigos de accesión de las variantes y su composición de bases.

5- Realiza el alineamiento de los genomas virales y visualiza el resultado de tu alineamiento en tu navegador. Muestra el código empleado para realizar lo anterior e incluye dos imágenes con el resultado del alineamiento, una de los primeros 150 nucleótidos y otra de los nucleótidos 500 al 650.

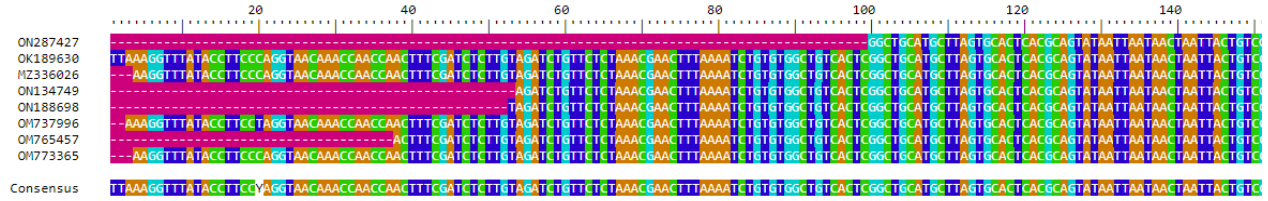

Figura 1.7 Imagen obtenida a partir de la IDE Rscript donde se plasma de forma grafica alineamiento de los genomas virales donde se nos muestran los primeros 150 nucleótidos de las secuencias variantes del virus Sars Cov-2.

A través de la **figura 1.7**, podemos observar que los primeros 150 nucleótidos de los genomas de las variantes de estudio contienen varios espacios en blanco cerca del inicio de sus secuencias. Sin embargo, mantienen un consenso extremadamente similar a lo largo de sus secuencias de nucleótidos.

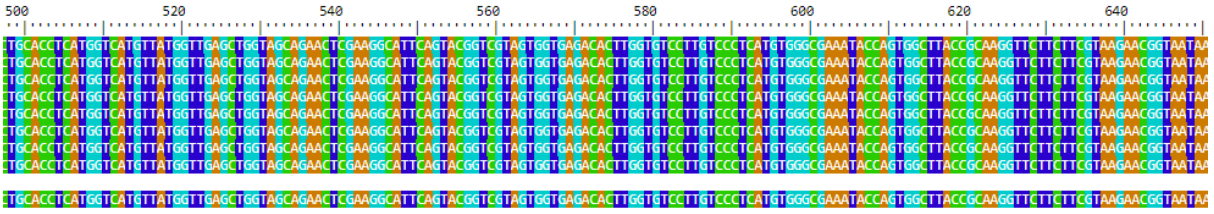

Figura 1.8 Imagen obtenida a partir de la IDE Rscript donde se plasma de forma grafica alineamiento de los genomas virales donde se nos muestran los de los 500-650 nucleótidos de las secuencias variantes del virus Sars Cov-2.

La **figura 1.8** nos muestra una secuencia de nucleótidos extremadamente uniforme entre todos los genomas de nuestras ocho variantes. Gracias a esto, podemos inferir que las mutaciones que crearon estas variantes se originaron en los primeros nucleótidos de sus secuencias.

6- Genera una matriz de distancia a partir de los genomas alineados. Crea una tabla en escala de grises en la que observes de manera visual el resultado de la matriz de distancia e inclúyela en tu reporte. Muestra el código empleado para obtener lo anterior e incluye la tabla que obtuviste.

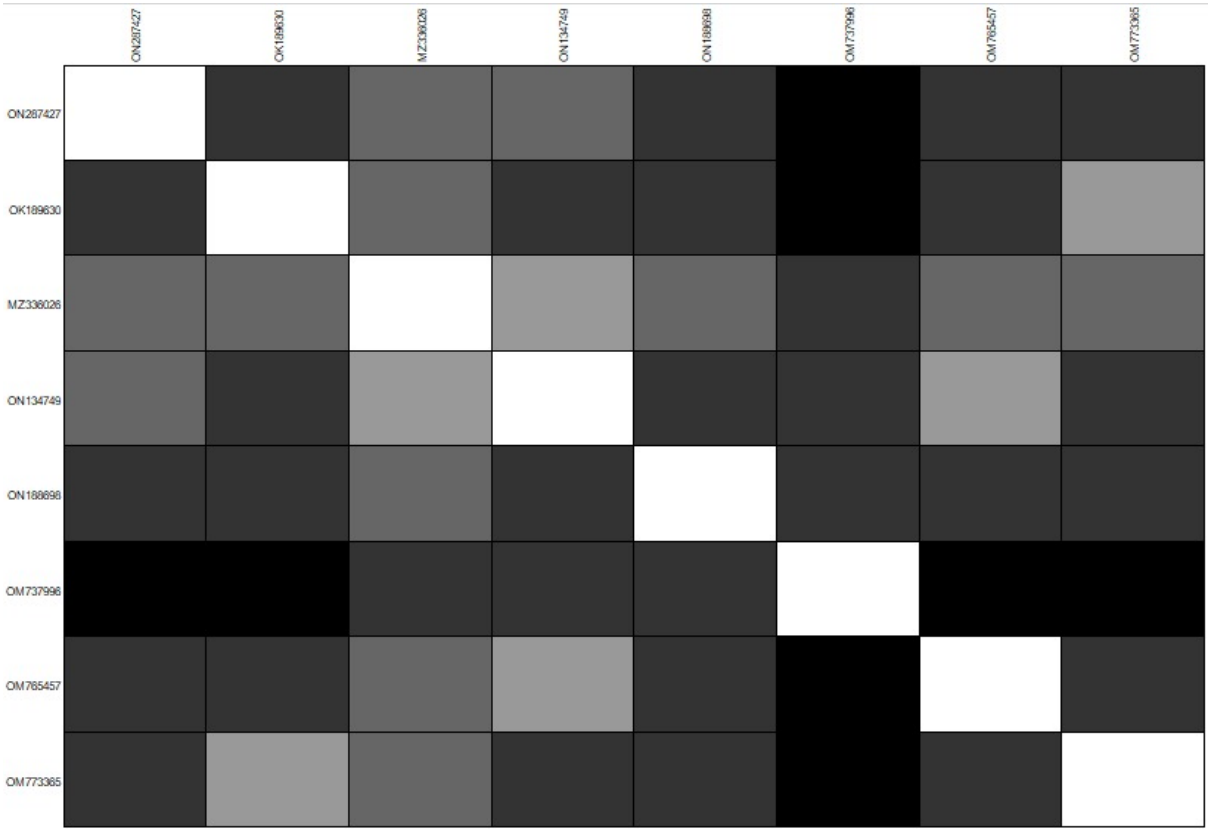

Figura 1.9 Imagen obtenida a partir de la IDE Rscript donde se plasma de forma grafica la tabla en escala de grises en la que se observa de manera visual el resultado de la matriz de distancia de los alineamientos de los genomas de las secuencias variantes del virus Sars Cov-2.

A través de la **figura 1.9,** podemos observar una representación gráfica en escala de grises de la matriz de distancia generada previamente. Los diferentes tonos de gris nos muestran las distancias entre los genomas de la matriz, con los cuadros blancos siendo una distancia de 0, es decir, representan la intersección de un mismo genoma. Los cuadros negros representan las intersecciones más lejanas entre genomas.

7-Construye un árbol filogenético a partir de la matriz de distancia obtenida e incluye en el árbol los números de accesión de los genomas utilizados, sus nombres o cualquier otra leyenda que te permita indicar la ubicación de ellos en el árbol.

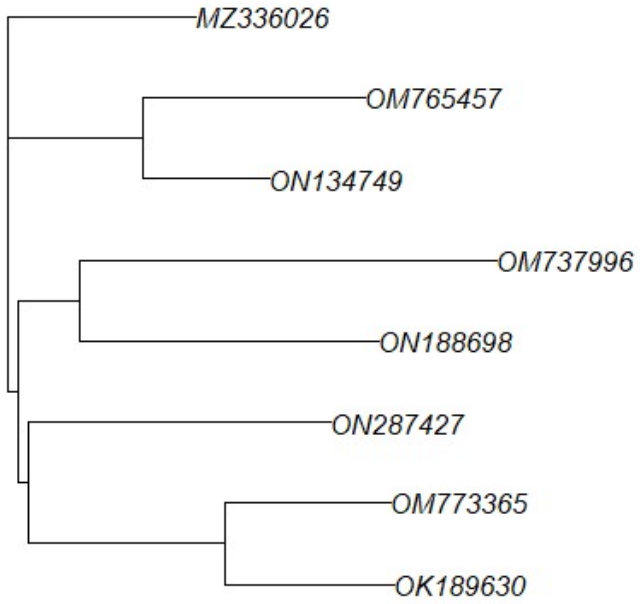

Figura 1.10 Árbol filogenético obtenido a partir de la IDE Rscript donde se plasma de forma grafica el arbol a partir de la matriz de distancia obtenida de las secuencias variantes del virus Sars Cov-2.

En la **figura 1.10**, podemos observar las relaciones ancestrales y familiares que comparten las diferentes variantes de SARS-CoV-2 que se examinaron en este proyecto. La gran mayoría de estas variantes comparten su ancestría inmediata con solo una otra variante.

Conclusiones

Gracias a la serie de herramientas y competencias computacionales orientadas en el lenguaje R y del contexto biológico que aprendimos y desarrollamos a lo largo del curso en cuestión del manejo de bases de datos biológicos en las distintas actividades y en esta situación problema, pudimos establecer una serie de obtención de datos de valor que nos ayudarán a establecer un análisis congruente y relevante a los cuestionamientos planteados a lo largo de este Proyecto Integrador.

Dado que nuestra situación problema se enfocaba en el análisis de variantes del virus SARS-CoV-2 causante del COVID-19 y que eran provenientes de distintas partes del mundo, pudimos estudiar gracias a las competencias y herramientas anteriormente descritas, aspectos clave como la longitud del genoma, la concentración de GC, y la distribución de sus bases (A,G,C,T) de ADN del genoma respectivo de cada variante del virus, en donde pudimos establecer conclusiones como: Que todas las variantes de este virus comparten un factor común en cuestión de las magnitudes y proporciones en su distribución de bases, pareciendo casi imperceptibles a simple vista y variando en un rango muy limitado a nivel decimal; otra conclusión fue el dato destacable en el que observamos que según la procedencia o país de origen de la variante del virus esta variaba en cuestión de la longitud del genoma, en el que obtuvimos el siguiente orden de mayor a menor: India, Sudáfrica, Oceanía, Estados Unidos y finalmente Francia, y siendo muy similar la longitud del genoma entre las variantes del mismo país de origen a pesar de ser otro tipo de variante.

A pesar de que estas variantes surgieron y se propagaron en regiones completamente distintas del mundo, las longitudes y contenidos de GC de sus genomas fueron bastante similares. La variación en sus composiciones de bases ocurrió principalmente en los primeros 0 a 200 nucleótidos, mientras que las diferencias entre sus nucleótidos después de este punto fueron casi imperceptibles. Como uno se podría imaginar, las distancias entre estas variantes son, por la mayor parte, mínimas. Esto se puede apreciar en la gráfica de la **figura 1.9.**

Referencias Bibliográficas

- Vera, F., Soulier Faure, M., Adler, V., Rojas, F., & Acevedo, P. (2020). *Ciudades Sostenibles*. BID. https://blogs.iadb.org/ciudades-sostenibles/es/pandemia-coronavirus-covid19-asentamientos-barrios-informales-medidas-emergencia-recuperacion/
- PubMed Central. (2021). *Variantes de SARS-CoV-2, una historia todavía inacabada*. NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8275477/
- PubMed. (s. f.). Spotlight on avian coronaviruses. National Library of Medicine. Recuperado 8 de mayo de 2022, de https://pubmed.ncbi.nlm.nih.gov/32374218/
- Jones, L. V. (2022, 2 marzo). Canine Coronavirus. PetMD. Recuperado 8 de mayo de 2022, de https://www.petmd.com/dog/conditions/digestive/c_dg_canine_coronavirus_infection
- The Coronaviruses of Animals and Birds: Their Zoonosis, Vaccines, and Models for SARS-CoV and SARS-CoV2. (2020, 24 septiembre). Frontiers. Recuperado 8 de mayo de 2022, de https://www.frontiersin.org/articles/10.3389/fvets.2020.582287/full
- Reina, J. (2020). El SARS-CoV-2, una nueva zoonosis pandémica que amenaza al mundo. NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270469/#:~:text=Los%20animales%20que%20s%C3%AD%20han,%2C%2017%2C%2018%2C%2019.

Video de nuestra presentación: https://youtu.be/4c0DXZIdZT8