Evidencia 2 | Análisis Final

Análisis geográfico de las secuencias de SARS-CoV-2: una comparación global de la diversidad genética del virus

Análisis de biología Computacional

Profesor: Heriberto García Coronado

Héctor Alán Gutiérrez Gálvez - A01253031

Camila Guadalupe Rodríguez Martínez - A01253767

Campus Sonora Norte

5 de Mayo, 2023

Tecnológico de Monterrey

Título del estudio: Análisis geográfico de las secuencias de SARS-CoV-2: una comparación global de la diversidad genética del virus.

Introducción

Un virus es un agente parasitario microscópico y acelular capaz de reproducirse en el interior de una célula hospedadora, por lo general ocasionado daños. Los virus son capaces de infectar a animales, plantas, bacterias e incluso otros virus llamados virófagos. (Equipo editorial, Etecé, 2021). En este trabajo, nos enfocaremos en el virus SARS-COV-2 (también conocido como covid-19), el causante de la pandemia mundial iniciada en el 2020. Es una enfermedad que tiene una gran afectación en el pulmón, debido a la forma en que se propaga y se replica en el cuerpo humano, además de que es extremadamente contagiosa. Pertenece a la familia Coronaviridae y se divide en cuatro géneros: alfa, beta, gamma y delta. En la actualidad se han descubierto siete coronavirus en humanos que están encuadrados en los géneros alfa y beta. (Fernández-Pérez et al., 2021).

En el mundo hay 676,609,955 casos, han habido 6,881,955 muertes y existen 13,338,833,198 personas vacunadas. En México se han reportado 7,483,44 casos y 333,188 muertes mientras que 225,063,079 personas ya fueron vacunadas. En el estado de Sonora fueron reportados 201,304 casos y 10,349 muertes. En Hermosillo hay un total de 82 mil casos y 3 mil defunciones. (*COVID-19 Map - Johns Hopkins Coronavirus Resource Center*, n.d.).

La primera variante del virus SARS-CoV-2 que se propagó a nivel mundial fue la variante D614G. Esta variante fue identificada por primera vez en Europa en febrero de 2020 y luego se extendió rápidamente a todo el mundo, convirtiéndose en la cepa dominante del virus a nivel global. (Aguilar-Gamboa et al., 2021).

Las variantes de SARS-CoV-2 más conocidas son la variante del Reino Unido (Alpha), Sudáfrica (Beta), Brasil (Gamma) y la de India (Delta). Otras variantes del virus, que existen en otra regiones del mundo son las siguientes: Epsilon también conocida como variante de California, Eta identificada por primera vez en Diciembre del 2020 en Nigeria y la variante Kappa identificada por primera vez India en Diciembre del 2020. (*Enfermedad Del Coronavirus 2019 (COVID-19)*, 2020). De igual manera, se han identificado varios coronavirus en otras especies animales que comparten similitudes genéticas con el SARS-CoV-2. A continuación, se presentan algunas de las especies animales en las que se han identificado coronavirus similares al SARS-CoV-2:

- Pangolines: se ha informado que los coronavirus similares al SARS-CoV-2 se han encontrado en pangolines, específicamente en la especie de pangolín de escamas largas (Manis pentadactyla) y la especie de pangolín de escamas chinas (Manis javanica). Estos coronavirus tienen una similitud genética del 85% al 92% con el SARS-CoV-2 (Lam, et al., 2020).

- Murciélagos: se ha informado que los coronavirus similares al SARS-CoV-2 se han encontrado en varias especies de murciélagos, incluidos los murciélagos herradura chinos (Rhinolophus spp.) y los murciélagos de herradura de alas largas (Miniopterus spp.). Estos coronavirus tienen una similitud genética del 96% al 97% con el SARS-CoV-2 (Zhou, et al., 2020).

La biología computacional es una disciplina interdisciplinaria que utiliza herramientas computacionales y técnicas de análisis de datos para abordar problemas biológicos complejos. Se basa en la integración de la biología molecular, la bioquímica, la genética, la estadística y la informática para desarrollar y aplicar métodos computacionales para analizar y comprender grandes conjuntos de datos biológicos. (*Diccionario De Cáncer Del NCI*, n.d.-b)

En resumidas cuentas el análisis de genomas virales mediante herramientas de biología computacional es una herramienta muy importante si se trata del estudio y control de enfermedades virales y este uso puede tener un gran impacto para la prevención y tratamiento de enfermedades infecciosas.

Objetivo

El fin de la realización de este estudio es lograr una comparación global de la diversidad genética del SARS-COV-2. Nuestra hipótesis es que en los diferentes países alrededor del mundo las características originales de las secuencias del SARS-COV-2 cambiarán cierto porcentaje. Esto se puede deber a muchos factores, por ejemplo el clima y la zona geográfica. Por lo tanto las preguntas de investigación que se busca responder son las siguientes: ¿Son muy diferentes las variantes entre cada país? ¿Es diferente el SARS-CoV-2 entre los diferentes países?

Tecnológico de Monterrey

Métodos y resultados

1. **Obtén las secuencias de los genomas de los virus elegidos según la investigación que hayas decidido realizar. Para ello utiliza la función read.GenBank para obtener los genomas directamente del NCBI desde R Studio.**

Código:

```
library(Biostrings)
library(ade4)
library(seqinr)
library(adegenet)
library(ape)


setwd("C:/Users/alang/Documents/Tec/Tareas/Analisis de Biologia Computacional")


corona_virus <- c("NC_045512", "OP435368", "OQ918256", "BS007312", "OQ913932", "OP848485", "ON291271", "MT994849",
                  "OK096766", "MW466791")


virus_sequences <- read.GenBank(corona_virus)
virus_sequences


write.dna(virus_sequences,file = "coronavirus_seqs.fasta",format = "fasta")

virus_seq_not_align <- readDNAStringSet("coronavirus_seqs.fasta", format = "fasta")
class(virus_seq_not_align)

virus_seq_not_align
```

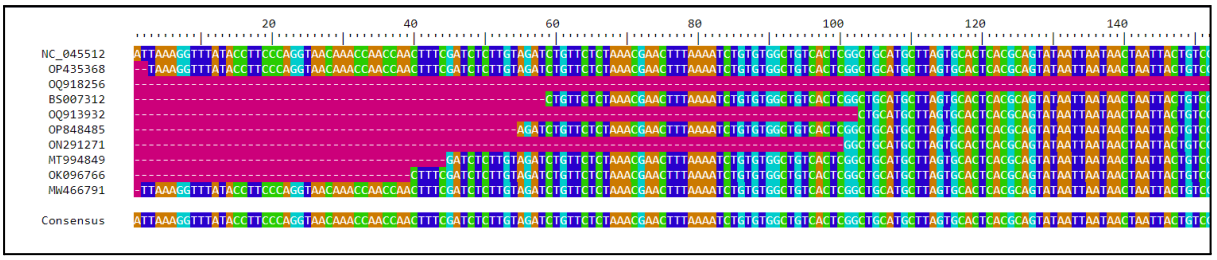
2. **Realiza el alineamiento de los genomas virales y visualiza el resultado de tu alineamiento en tu navegador web. Muestra el código empleado para realizar lo anterior e incluye dos imágenes con el resultado del alineamiento, una de los primeros 150 nucleótidos y otra de los nucleótidos 500 al 650.**

Código:

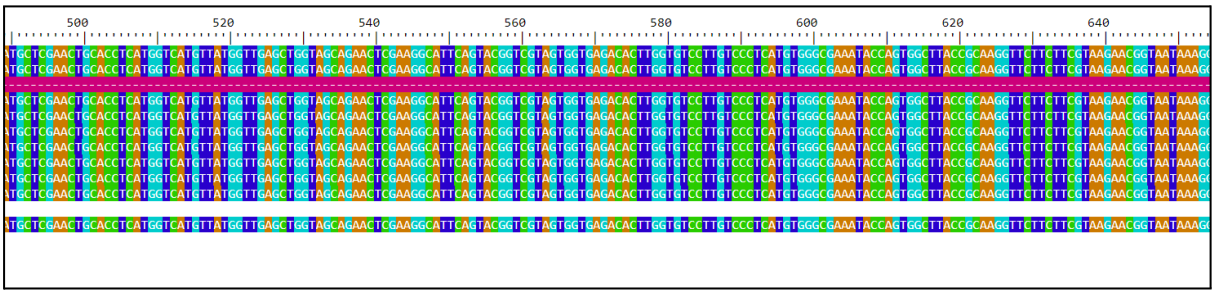
```r
library(DECIPHER)


# Alineamiento de las primeras 150 posiciones
virus_seq_not_align_150 <- virus_seq_not_align[,1:150]
virus_seq_not_align_150 <- OrientNucleotides(virus_seq_not_align_150)
virus_seq_align_150 <- AlignSeqs(virus_seq_not_align_150)

# Alineamiento de los nucleótidos 500 al 650
virus_seq_not_align_500_650 <- virus_seq_not_align[,500:650]
virus_seq_not_align_500_650 <- OrientNucleotides(virus_seq_not_align_500_650)
virus_seq_align_500_650 <- AlignSeqs(virus_seq_not_align_500_650)


BrowseSeqs(virus_seq_align_150)
BrowseSeqs(virus_seq_align_500_650)

```

Resultado en consola:

Alineamiento de las primeras 150 posiciones:



Alineamiento de los nucleótidos 500 - 650:



3. **Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica (de 6 a 12 renglones).**

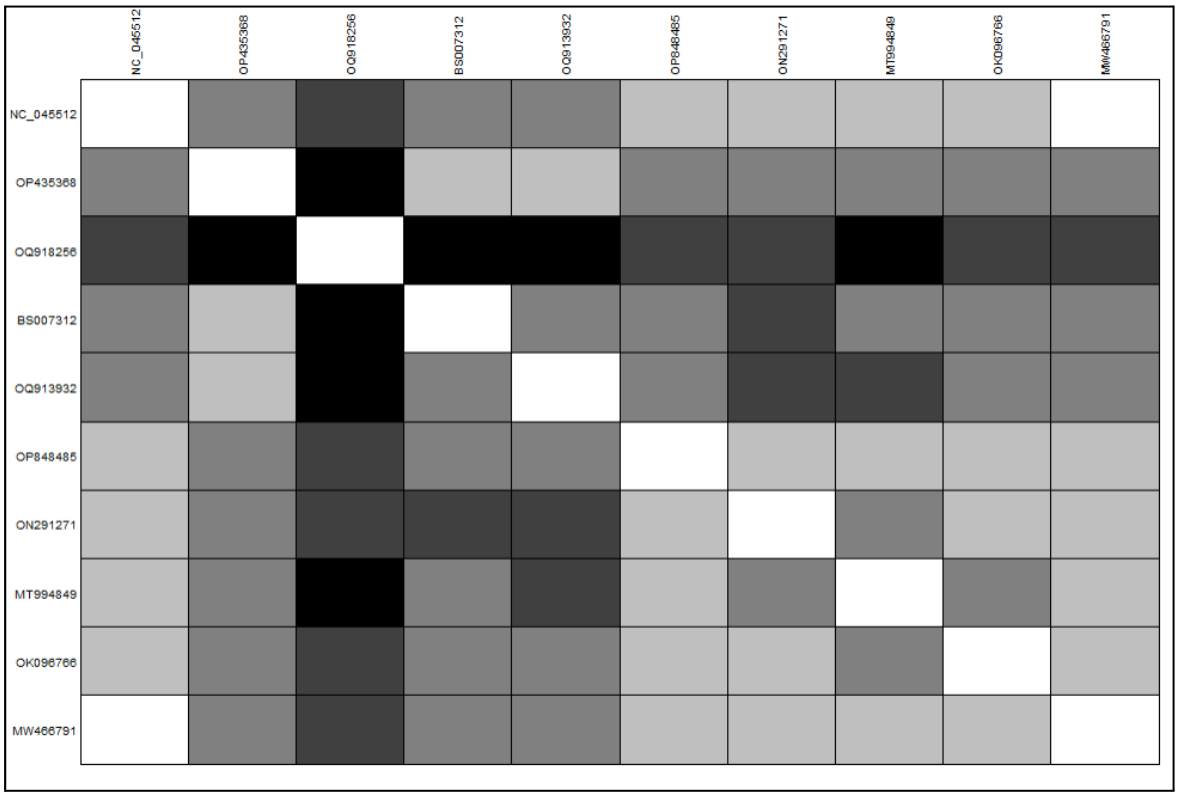
Con este proceso de alineamiento de los genomas, podemos observar claramente la comparación entre las secuencias, ya que indica donde coinciden sus bases nitrogenadas.
Desde el punto de vista biológico se puede observar que los 10 genomas de SARS-COV-2 son muy parecidos. Sin embargo, también muestra que los genomas originarios de diferentes países alrededor del mundo sufren de alteraciones, esto lo podemos ver por ejemplo en el caso del alineamiento de las primeras 150 posiciones, donde existen más diferencias entre ellos. Por otro lado, en la gráfica de alineamiento de los nucleótidos 500 - 650 tienen bastante parecido, ya que sus bases coinciden entre ellas.

4. **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 matriz de distancia y la tabla que obtuviste.**

Código:

```
55
56
57 writeXStringSet(virus_seq_align_150, file = "coronavirus_seq_align_150.fasta")
58 writeXStringSet(virus_seq_align_500_650, file = "coronavirus_seq_align_500_650.fasta")
59
60
61 virus_aligned_150 <- read.alignment("coronavirus_seq_align_150.fasta", format = "fasta")
62 virus_aligned_500_650 <- read.alignment("coronavirus_seq_align_500_650.fasta", format = "fasta")
63
64
65 matriz_distancia_150 <- dist.alignment(virus_aligned_150, matrix = "similarity")
66 matriz_distancia_500_650 <- dist.alignment(virus_aligned_500_650, matrix = "similarity")
67 as.data.frame(as.matrix(matriz_distancia_150))
68 as.data.frame(as.matrix(matriz_distancia_500_650))
69
70 tablas_grises_150 <- as.data.frame(as.matrix(matriz_distancia_150))
71 tablas_grises_500_650 <- as.data.frame(as.matrix(matriz_distancia_500_650))
72
73 table.paint(tablas_grises_150, cleg = 0, clabel.row = .5, clabel.col = .5)
74 table.paint(tablas_grises_500_650, cleg = 0, clabel.row = .5, clabel.col = .5)
75
```

Resultado en consola:



5. **Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica (de 6 a 12 renglones).**

Desde el punto biológico, esta tabla de colores representa cada valor en la matriz de distancia o similitud como un cuadrado coloreado en la tabla, en este caso entre más oscuro esté el recuadro existe mayor diferencia, por lo tanto si el recuadro es blanco será muy parecido (en algunos casos, completamente igual). Por ejemplo, podemos dar cuenta que el OQ913932 (Estados Unidos) y el OP848485 (Australia) no muy parecidos en base a la tabla y el virus MW466791 (Corea del Sur) y el NC_045512 (China) son muy similares. Con la información

anterior, concluimos que los genomas que se encuentran geográficamente más cercanos, tendrán más parecido. Mientras que si los países se encuentran más alejados, los genomas tendrán menos parecido uno con el otro.

Al analizar la tabla se puede observar que algunos genomas tienen menos variación con otros, lo cual indica que existen variaciones en los genomas de los diferentes SARS-COV-2, dependiendo del país en donde se encuentren. Esto nos puede ayudar mucho para el análisis y detección de muchos virus en el mundo y encontrar cura de ciertos virus mediante la comparación.

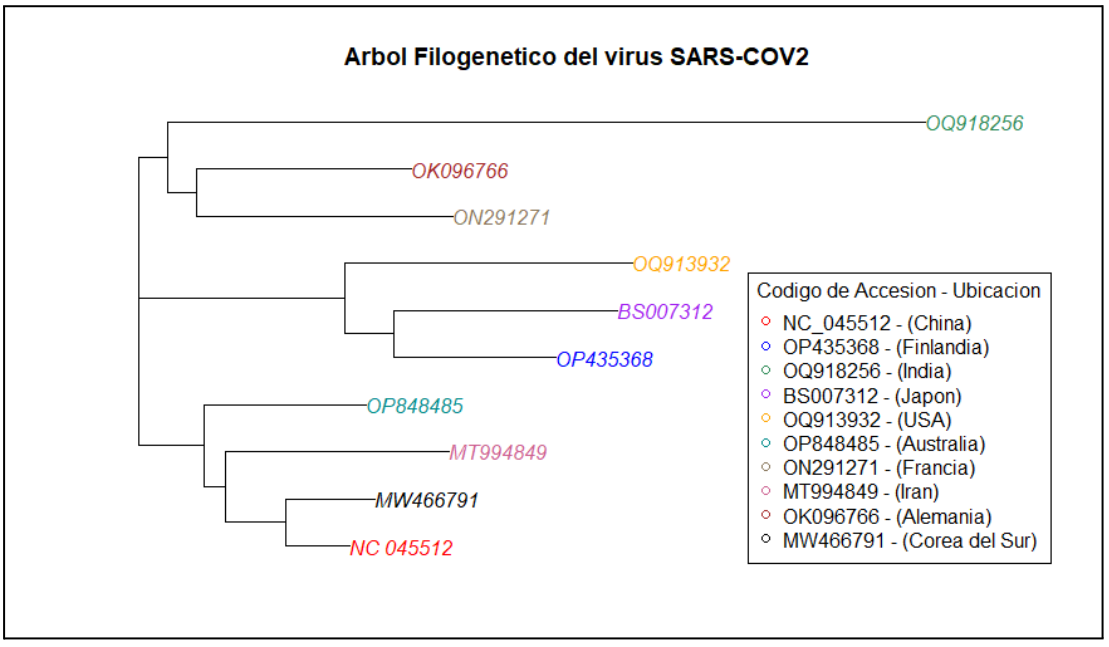
6. **Construye un árbol filogenético a partir de la matriz de distancia obtenida e incluye en el árbol los números de accesión de los genomas utilizados, sus nombres comunes o cualquier otra leyenda que te permita indicar la ubicación de cada virus en el árbol. Muestra el código empleado para realizar lo anterior e incluye la imagen del árbol filogenético.**

Código:

```r
library(phytools)
library(maps)
library(viridis)
library(viridisLite)
library(ggtree)
library(ggplot2)

virus_tree <- nj(matriz_distancia_150)
virus_tree2 <- nj(matriz_distancia_500_650)


virus_colors <- c("red", "blue", "#2E8B57", "purple","orange", "#008B8B",
                  "#8B795E", "#CD6090", "brown", "black")
virus_tree <- ladderize(virus_tree)


plot(virus_tree, main = "Arbol Filogenetico del virus SARS-COV2", tip.color=virus_colors)

plot_virus_filogenia <- ggtree(virus_tree) +
  geom_tiplab(aes(color=virus_id)) +
  ggtitle("Analisis Filogenetico de Virus")

# Agregar virus_id a tip_dates
tip_dates$virus_id <- virus_id

```

```r
# Definir paleta de colores
colores_virus <- c("NC_045512" = "red", "OP435368" = "blue", "OQ918256" = "#2E8B57",
                   "BS007312" = "purple", "OQ913932" = "orange", "OP848485" = "#008B8B",
                   "ON291271" = "#8B795E", "MT994849" = "#CD6090", "OK096766" = "brown",
                   "MW466791" = "black")

# Agregar leyenda de colores
plot_virus_filogenia <- plot_virus_filogenia +
  scale_color_manual(values = colores_virus) +
  guides(color=guide_legend(title="Codigo de Accesion y Ubicacion")) +
  theme(legend.position="bottomright")

# Agregar leyenda con nombres y ubicaciones de virus
legend("bottomright", legend = paste(tip_dates$tips, " - (", tip_dates$ubicacion, ")", sep = ""),
       pch = 1, col = colores_virus[virus_id], title="Codigo de Accesion - Ubicacion")

```

Resultado en consola:



Tecnológico de Monterrey

7. **Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica (de 6 a 12 renglones).**

Este árbol filogenético es una representación gráfica que nos muestra las relaciones evolutivas entre 10 virus tipo SARS-COV2. Fue construido mediante el análisis de similitudes y diferencias en características como lo es la ubicación en el que se manifestó el virus. Las ramas representan la divergencia de diferentes cepas de virus a lo largo del tiempo y los nodos representan el punto donde ocurrieron divergencias en la evolución de los virus, por ejemplo el virus con código de accesión OK96766, detectado en Alemania, y el genoma originario de Francia ON291271 comparten un ancestro en común, ya que sus ramas originaron en un mismo nodo. Con la información anterior, se puede concluir que en ciertos casos los lugares cercanos tendrán ancestros en común, debido a las variables que se toman en cuenta, como lo puede ser el clima y el estilo de vida de las personas. Estos árboles son muy útiles para conocer la evolución de características específicas en diferentes linajes de organismos.

Conclusión

En conclusión podemos deducir que las variantes en cada país son bastantes similares, unos genomas mantienen más diferencia pero en general la mayoría tiene características en común, esto se puede deber a la forma de vida que se tiene en los diferentes países. Por otro lado el SARS-COV2 varía si los países están más alejados geográficamente, esto lo podemos ver en las gráficas que realizamos anteriormente, ya que entre los 10 virus muestran un gran grado de similitud entre más cercanos se encuentren.

En resumen el uso de gráficas y tablas es muy útil para el análisis de genomas virales tales como el famoso SARS-COV2 que con la ayuda de lo mencionado anteriormente, puede ser de una gran ayuda para tener un mejor conocimiento de los virus y poder evitar tragedias, como lo fue de pandemia de Covid-19.

Tecnológico de Monterrey

LINK DEL VIDEO: https://youtu.be/Xu5f_dXRw4w

Referencias:

- "Virus en Biología". Autor: Equipo editorial, Etecé. De: Argentina. Para: *Concepto.de*. Disponible en: https://concepto.de/virus-en-biologia/. Última edición: 5 de agosto de 2021. Consultado: 02 de mayo de 2023 Fuente: https://concepto.de/virus-en-biologia/#ixzz80b68fAYF

- Fernández-Pérez, G. C., Miranda, M. O., Fernández-Rodríguez, P., Casares, M. V., De La Calle, M. C., López, Á. F., Blanco, M., & Cuchat, J. M. O. (2021). SARS-CoV-2: cómo es, cómo actúa y cómo se expresa en la imagen. *Radiología*, *63*(2), 115–126. https://doi.org/10.1016/j.rx.2020.10.006

- *COVID-19 Map - Johns Hopkins Coronavirus Resource Center*. (n.d.). Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html

- *Situación de COVID-19 : Secretaría de Salud – COVID 19*. (n.d.). http://covid19.saludsonora.gob.mx/?page_id=2339

- Aguilar-Gamboa, F. R., Suclupe-Campos, D. O., Vega-Fernández, J. A., & Silva-Diaz, H. (2021). Diversidad genómica en SARS-CoV-2: Mutaciones y variantes. *REVISTA DEL CUERPO MÉDICO HOSPITAL NACIONAL ALMANZOR AGUINAGA ASENJO*, *14*, 572–582. https://doi.org/10.35434/rcmhnaaa

- *Enfermedad del coronavirus 2019 (COVID-19)*. (2020, February 11). Centers for Disease Control and Prevention. https://espanol.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html

- Lam, T. T., Jia, N., Zhang, Y. W., Shum, M. H., Jiang, J. F., Zhu, H. C., ... & Holmes, E. C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature, 583(7815), 282-285.

- Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579(7798), 270-273.

- *Diccionario de cáncer del NCI*. (n.d.-b). Instituto Nacional Del Cáncer. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/biologia-computacional

- Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nature Medicine, 26(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9