Proyecto.

Integrantes :Emiliano Dominguez Cruz.
Juan Antonio Jasso Oviedo.
Miguel Angel Liera Montaño.

14 de Diciembre de 2022

1 Introducción.

Los Coronavirus son un grupo de virus de RNA que causa infecciones o enfermedades respiratorias y digestivas en rango leve a leta en distintos mamíferos y aves. Se dividen en los generos Alphacoronavirus, Betacoronavirus, Gammacoronavirus y Deltacoronavirus.

Son estructuras grandes, casi esféricas con protuberancias en la superficie. Que al ser observadas más detenidamente hacen que el virus de la forma de un "halo" de luz o corona de luz, de ahí su nombre. Su estructura esta conformada por la envoltura y la nucleocápside. Los coronavirus contienen un RNA monocatenario de sentido positivo que van de los 26.4 a 31.7 kilobases. El genoma posee una tiene una tapa metilada en 5 y una cola poliadenilada en 3.

Existen diferentes teorias sobre cómo el coronavirus se transmitió al ser humano inicialmente. En principio, como cualquier virus que salta de huespedes de diferentes especies, este presenta una serie de cambios significativos hablando evolutivamente para lograr infectar otros organismos. En particular, el coronavirus, incluso antes de la transmisión a humanos poseía una capacidad notable para infectar a otros organismos.

Una vez transmitido al ser humano, se sabe que las variantes del virus más resistentes al sistema inmunologico del humano fueron aquellas que prevalecieron, gracias a un mayor número de infectados.

Las mutaciones en los virus son constantes y la importancia de estudiar la presedencia de estos cambios y qué tanto se diferencian cada tipo de virus es poder evaluar las tendencias del virus y poder generar nuevas defensas que contrarresten las nuevas variantes. Por ello, se mostrará a continuación un estudio sobre las diferencias de los distintos tipos del coronavirus en diferentes organismos, a fin de conocer las similitudes entre estas.

1.1 Pregunta de investigación.

¿Qué tanto divergen los genomas de los coronavirus de distintas especies entre si?

1.2 Hipótesis.

Coronavirus de aves contra el de mamiferos es distinto o incluso entre mamíferos puede cambiar.

1.3 Objetivo General.

Determinar el grado de similitud entre los genomas de distintos coronavirus, que afectan a distintos organismos(mamíferos,aves).

1.4 Obejtivos Particulares.

- Realizar conteo de GC.
- Indice de similitud entre las cadenas.
- Alineamiento y aplicación de algoritmo Needleman-Wuntch.

2 Resultados.

2.1 Conteo de GC.

Genoma	Citocinas	Guaninas
Musaraña de la india	23	63
Rata	53	64
Erizo	55	93
Camello	4591	5930
Ganso Canadiense	4738	6230
Humano	4549	5903
Vison	4690	6155
Murcielago Rosteo	6407	7257

Analizando la tabla podemos notar que las diferencias entre la cantidad de Citocinas no es tanta, donde los genomas completos(Camello,Ganso Canadiense,Humano,Vison y Murcielago Rosteo) están la mayoría en el rango de $4500 - 4780$ Con la excepción del murcielago que tiene una mayor cantidad de Citocinas. Además en la parte de Guaninas, la diferencia es un poco más grande pero de nuevo no son abismales las diferencia con la mayoría entre $5900 - 6250$ y de nuevo el camello tiene la mayor cantidad de Guaninas.

Las secuencias parciales que encontramos su rango de diferencia es incluso más pequeño que en el de los Genomas completos. Por ejemplo, el de Rata y erizo solo difieren en 2 unidades de Citocina pero el de Musaraña y Rata solo difieren en una unidad de Guanina.

Como vimos en la práctica 3 el contenido de GC dentro de un genoma es relevante conocerlo ya que puede se un indicador del tamaño del genoma y sobre las estructuras cromosomales holocentricas de los organismos.

Particularmente en los virus-RNA como el coronavirus, de acuerdo al articulo de NCBI listado en las fuentes como *Composition bias and ...*, el contenido de *GC* puede ser un factor importante en el sesgo de uso de codones, que afecta la efficiencia de la expresión. También afecta en su polaridad del genoma, ya que los "positive-stranded" virus tienen un alto contenido de *GC*, significativamente mayor a los "negative-strande". Y el coronavirus es uno positivo.

2.2 Alineamiento de Cadenas y Needleman-Wuntch.

Intentamos utilizar el algoritmo de alineacion Needleman-Wuntch y el sistema de puntuacion que programamos en una practica pasada para hacer la comparacion entre los genomas, pero encontramos que no poseiamos la capacidad de procesar cadenas tan grandes por lo que decidimos limitar la cantidad de nucleotidos que considerariamos en el alineamiento a solo 1600, y utilizar colaboratory de google para tener mas poder de procesamiento

Cadenas a comparar	Puntuacion	Cadenas a comparar	Puntuacion
Humano vs Camello	12955	Camello vs Vison	1000
Humano vs Ganso Canadiense	-1592	Camello vs Murcielago Rosteo	-1954
Humano vs Vison	1104	Ganso Canadiense vs Vison	-1141
Humano vs Murcielago Rosteo	-1965	Ganso Canadiense vs Murcielago Rosteo	-2695
Camello vs Ganso Canadiense	-1496	Vison vs Murcielago Rosteo	-2336

Podemos observar que los genomas de Humano y Camello presentan no solo el puntaje mas alto tambien el puntaje mas alejado del resto, mientras que los demas puntajes oscilan entre -2700 y 1200, (una distancia de 3900 puntos) el puntaje de Humano vs Camello se separa por mas de 10000 puntos del mas cercano. Tambien podemos observar el Murcielago Rosteo es el que presenta peores puntajes, no alcanzando ni uno solo positivo, lo cual es algo que concuerda con nuestro conteo de GC donde era el que mas diferia

2.3 Obtención de indice de similitud entre cadenas.

A continuación una tabla donde veremos la distancia de el genoma del coronavirus humano contra otras secuencias. Esta distancia será siempre un entero y nos ayudará a determinar índices de similitud entre las secuencias o cadenas de las cuales ya sacamos el conteo de GC.

Esto lo hacemos con un algoritmo que proponemos donde, tomamos en cuenta que nuestras cadenas pueden no ser de la misma longitud(que evita la distancia de Hamming sea util) y donde ya que obtenemos la diferencia de longitudes podemos pasar a comparar los índices que sabemos podemos comparar entre las cadenas.

Nuestra distancia representa el indice que obtenemos de comparar las cadenas, el número de matchs se refiere a cuántas veces nuestras cadenas en la misma posición tienen la misma letra(Nucleotido) y la diferencia de longitudes nos ayuda a manejar justamente cuando las cadenas no son de la misma longitud(lo cual ocurre muy seguido).

Cadenas a comparar	Distancia.	Número de Matchs.	Diferencia de longitudes.
Humano vs Camello	17458	10280	79
Humano vs Ganso Canadiense	20642	8254	1237
Humano vs Visón	21813	7490	1644
Humano vs Murcielago Rosteo	23303	7236	2880

Nuestro algoritmo propuesto se encuentra en nuestro repositorio de git, con el nombre “distancia”.

Ahora, observando la tabla notamos que el animal con el genoma que se tuvo menor distancia es con el coronavirus del camello. Se tuvo la menor distancia y también el máximo número de matchs además de que la diferencia de tamaños fue de solo $79pb$.

Así, es más fácil que un humano se infecte con la “cepa” de coronavirus de camello que con el de un Murcielago Rosteo, que fue el que mayor distancio nos generó y además tuvo el menor número de matchs.

Sin embargo, cabe recalcar que con el resto de genomas la diferencia se mantiene en un rango de $20000 - 23000$. Por lo que, nuestra similitud se mantiene entre los distintos genomas.

3 Conclusión.

Como podemos ver a lo largo de nuestros objetivos particulares, las similitudes entre las cadenas se mantienen, claro con algunas excepciones. Además, los resultados son congruentes ya que primero por el alineamiento utilizando Needleman-Wuntch el mejor fue Camello vs humano, además en nuestro algoritmo de distancia(que fue una propuesta nuestra de algoritmo) encontramos que la

menor distancia entre las cadenas y además el mayor número de matchs fue justamente Camello vs humano.

Además, podemos notar que en general al analizar nuestros resultados las similitudes se mantienen desde el contenido de GC, que se mantienen habiendo siempre las excepciones; el alineamiento es igualmente comparable con la tabla de distancias o indices de similitud obtenidas.

Por lo tanto, nuestra hipotesis es correcta. La similitud entre los coronavirus es observarble tras realizar un análisis de ciertas caracteristicas de las cadenas, ya sea con genomas completos o parciales.

4 Fuentes.

Coronavirus. `https://en.wikipedia.org/wiki/Coronavirus`

Secuencias de genoma. `https://www.ncbi.nlm.nih.gov/`

Composition bias and genome polarity of RNA-viruses. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7114242/`

SARS-CoV-2 jumped from bats to humans without much change, study finds `https://www.sciencedaily.com/releases/2021/03/210312155448.htm`