Práctica 5 - Bioinformática
Álvaro Romeo - 715810
21/04/2021

1.Introducción

El objetivo de esta práctica es analizar la variabilidad de SARS-CoV-2 sobre un conjunto de datos proporcionados y calcular además el árbol filogenético correspondiente que explique la evolución del virus.

Los datos proporcionados consisten en aproximadamente unas 400 secuencias de Cataluña (en este caso del reparto catHC2), que abarcan más o menos de entre septiembre de 2020 y enero de 2021, junto con la secuencia de referencia de Wuhan y una secuencia de la variante B.1.1.7 o cepa británica.

2.Multi Alineamiento

Para la realización del multi alineamiento se ha utilizado MAFFT, principalmente por el coste temporal para la realización de este. Se han probado distintas formas de realizar el alineamiento de las secuencias y cabe destacar que se ha removido una secuencia en concreto debido a que producía un gap de bastante longitud al final del alineamiento en el resto de secuencias, ya que en esas posiciones la secuencia en cuestión tenía una sucesión de N´s lo que producía dicho gap en el resto de secuencias.

Además, destacar también que los alineamientos producidos en mafft por las distintas pruebas realizadas eran bastante similares por lo que se ha escogido el más simple de ejecutar y en un tiempo razonable. El comando utilizado es el siguiente:

mafft --thread -1 --maxiterate 2 --adjustdirection input > output

Donde:

--thread -1: permite a mafft utilizar todos los cores disponibles en la máquina

--maxiterate 2: realiza un refinamiento iterativo sobre el método progresivo de FFT-NS-2 un número de 2 veces. Más iteraciones no implican unos mejores resultados ya que las mejoras en calidad se obtienen principalmente en las primeras iteraciones

--adjustdirection: permite ajustar la dirección del alineamiento de acuerdo con la primera secuencia (que es la de referencia) lo que es útil para que el alineamiento tome como guía la secuencia de referencia que es lo que se quiere.

3.Detección de mutaciones y errores de secuenciación

Una vez obtenido el alineamiento se ha implementado el algoritmo en dos fases para calcular la medida de conservación para cada posición del alineamiento.

Siguiendo el algoritmo, en una primera fase se calcula la frecuencia de cada nucleótido para cada posición del alineamiento, esto es recorriendo todo el fichero del alineamiento, secuencia por secuencia y almacenando las frecuencias para cada posición, para luego en la segunda fase calcular la medida de conservación para cada posición siguiendo el método de entropía propuesto. Cabe destacar que para calcular las frecuencias sólo se han tenido

en cuenta las 4 letras que indican un nucleótido (A,G,T,C) y el resto que pueden aparecer se han tenido en cuenta de la misma forma que si fuera un gap.

En primer lugar, se ha realizado un análisis de variabilidad siguiendo las zonas que tienen especial relevancia, tal y como se indica en el enunciado. La siguiente tabla muestra algunas medidas de interés recopiladas.

Zona	Max C(i)	Min C(i)>0	Media
Gen ORF1A Posiciones 266-13.483	0.690923 posición 5.169	0.0177033 posición 278	0.00140776
Proteína Spike Posiciones 21.563-25.384	0.524715 posición 22.226	0.0177033 posición 21617	0.0018038
Gen ORF3A Posiciones 25.393-26.220	0.250726 posición 25.846	0.0177033 posición 25.426	0.00259703
Gen N Posiciones 28.274-29.533	0.511458 posición 28.934	0.0177033 posición 28.302	0.00518023
Global	0.690923 posición 5.169	0.0177033 en varias	0.00166592

De la obtención de estas medidas tampoco se deduce demasiado, pero comparando con la gráfica de diversidad de https://nextstrain.org/ncov/europe?c=country algunas medidas como la máxima del Gen N (posición 28.934), no presenta un valor tan alto como el de la gráfica (0.65 para esa posición) pero sí que tiene un valor bastante alto de 0.511 de variabilidad, lo que se intuye que es un rango de posición donde suele haber cierta variabilidad y por tanto mutaciones. El rango de las 22.000 (gen ORF3A) también presenta grandes cambios de variabilidad, y en nuestro caso se obtiene 0.52 en la posición 22.226 lo que es un valor también bastante alto.

Además las posiciones obtenidas como mínimas son bastante bajas, pero comparando con el mismo gráfico ya mencionado anteriormente, esas posiciones presentan unos valores que no pasan el 0.05 de variabilidad, y por tanto en esas posiciones las mutaciones no son tan comunes tal y como se ha podido observar en los resultados obtenidos.

4.Construcción de una filogenia

Para la construcción de los árboles de filogenia se han generado varios árboles utilizando tanto RAxML como FastTree. Los modelos evolutivos en cada caso junto con los comandos ejecutados son los siguientes:

RAxML

1.Modelo evolutivo GTR con CAT:

./raxmlHPC-PTHREADS-AVX -s input.fasta -n tree.txt -m GTRCAT -p 111 -T 6

-De los cuales parámetros importantes son:

-m GTRCAT que indica que se está aplicando el modelo GTR con CAT.

-p 111 que especifica la semilla empleada para inferencias con parsimonia

2.Modelo evolutivo GTR con GAMMA

./raxmlHPC-PTHREADS-AVX -s input.fasta -n output.txt -m GTRGAMMA -p 111 -T 6

-De los cuales parámetros importantes son:

-m GTRGAMMA que indica que se está aplicando el modelo GTR con GAMMA, que es más lento que el anterior pero más preciso.

-p 111 que especifica la semilla empleada para inferencias con parsimonia

FastTree

1.Modelo evolutivo por defecto Jukes Cantor + CAT

./FastTree/FastTreeMP -nt input.fasta > output.txt

-Es el modelo evolutivo por defecto, que utiliza Jukes Cantor con CAT, y el parámetro *-nt* se utiliza para especificar que son secuencias de nucleótidos.

2.Modelo evolutivo GTR + CAT

./FastTree/FastTreeMP -gtr -nt input.fasta > output.txt

-Utiliza un modelo evolutivo GTR, especificado con el parámetro *-gtr* con CAT por defecto.

3.Modelo evolutivo Jukes Cantor + GAMMA

./FastTree/FastTreeMP -gamma -nt input.fasta > output.txt

-Es el modelo evolutivo por defecto, que utiliza Jukes Cantor y se añade *-gamma* para que se utilice gamma para el índice de similitud en vez de CAT.

4.Modelo evolutivo GTR + GAMMA

./FastTree/FastTreeMP -gamma -gtr -nt input.fasta > output.txt

-Utiliza un modelo evolutivo GTR, especificado con el parámetro *-gtr,* con GAMMA, especificado mediante el parámetro -gamma.

A continuación se muestra una comparativa de los resultados obtenidos con los distintos programas ejecutados para la construcción del árbol filogenético:

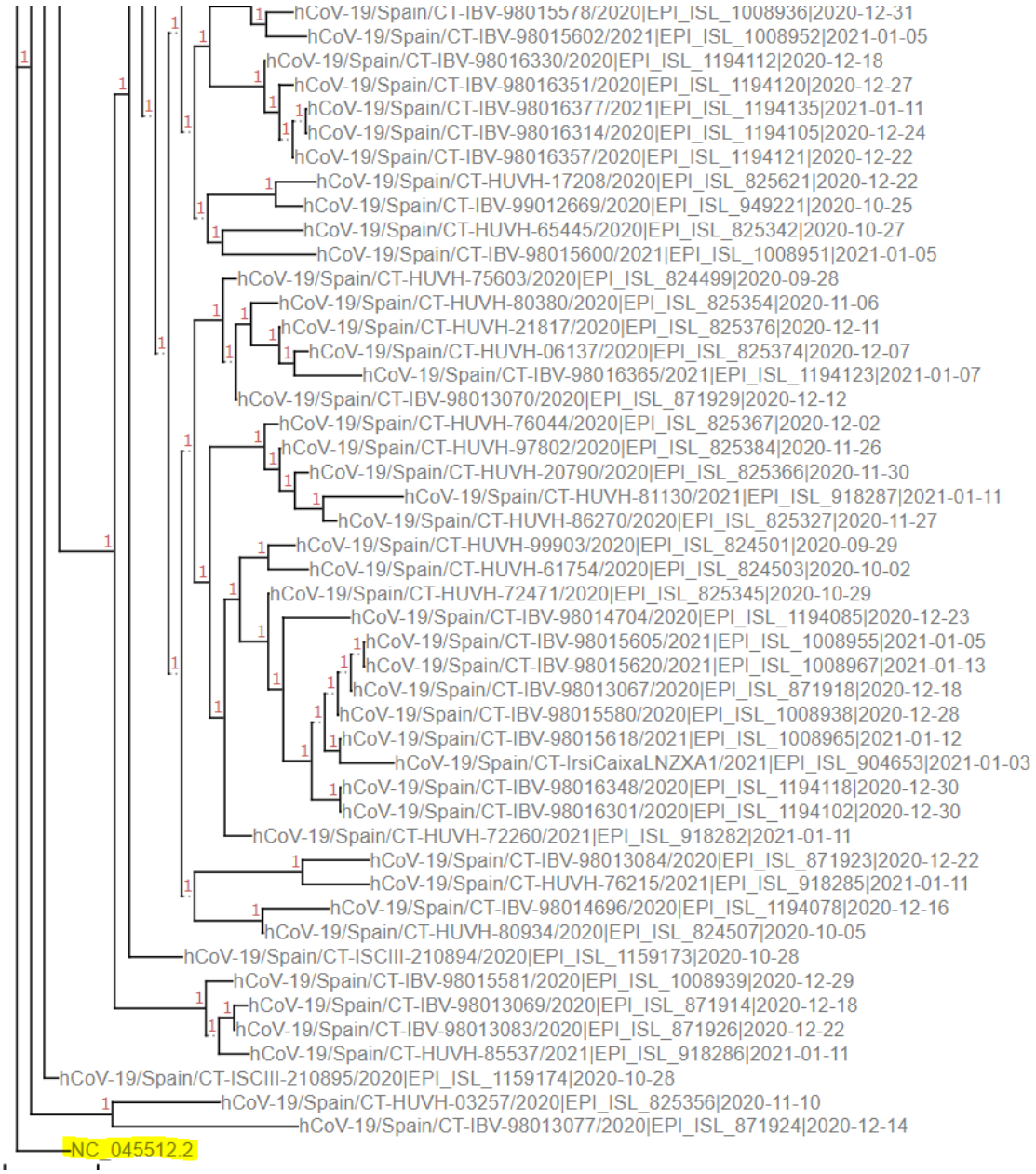
Programa	Modelo	Likelihood	Tiempo(segundos)
RAxML	GTR + CAT	-53744,097	120.82
RAxML	GTR + GAMMA	-53743,87	155.86
FastTree	JC + CAT	-64639,58	84.82
FastTree	GTR + CAT	-61586,186	312.59
FastTree	JC + GAMMA	-56556,802	146.49
FastTree	GTR + GAMMA	-55560,413	322.46

Como se puede observar, RAxML funciona mejor en ambos casos, tanto con GTR + CAT como con GTR + GAMMA, obteniendo unos índices de similitud mayores y en un tiempo menor, si los comparamos con los mismos modelos en FastTree.

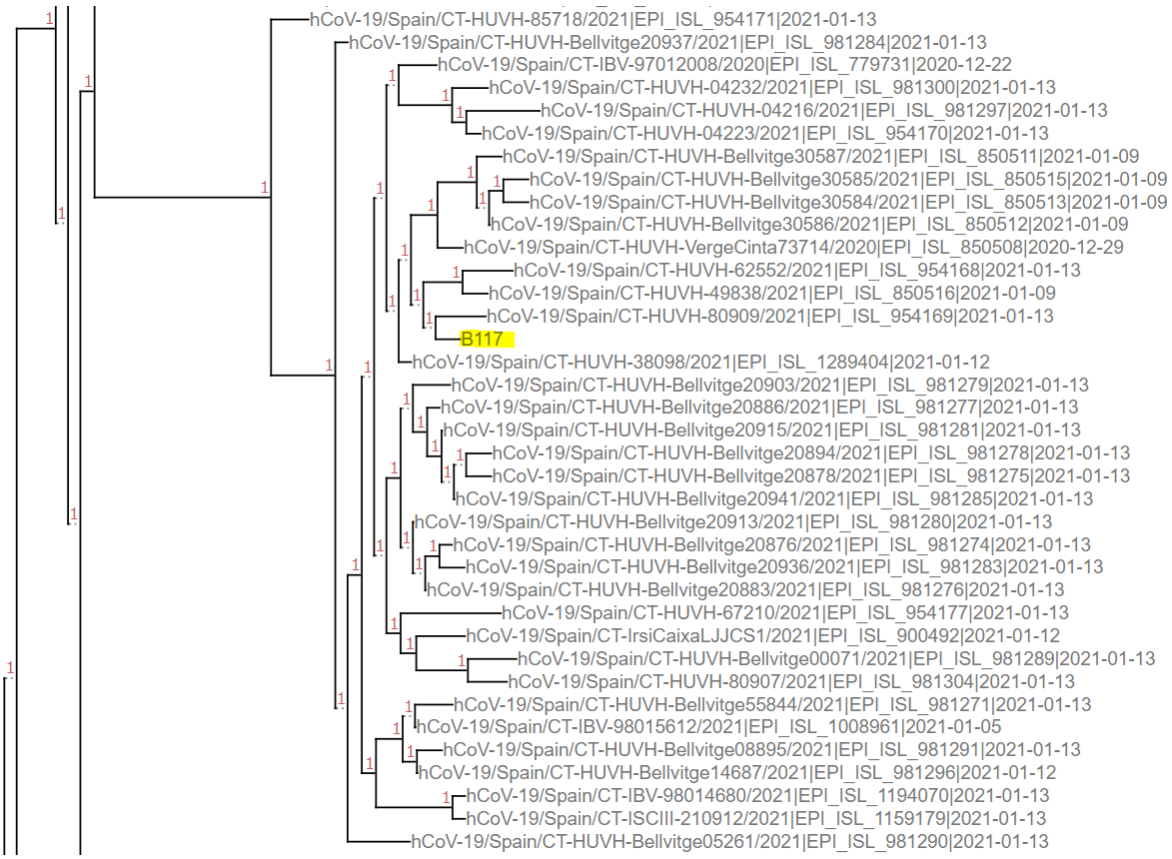
Por otro lado, dentro de FastTree, podemos observar que el modelo GTR obtiene un índice de similitud mejor en ambos casos, pero con el modelo de Jukes Cantor el tiempo se reduce a la mitad, obteniendo unos índices que no difieren tantísimo comparado con los otros dos teniendo en cuenta que se obtienen números muy grandes. Entre aplicar CAT o GAMMA la diferencia no es demasiado notoria, obteniendo unos resultados mejores con GAMMA

Esto concuerda con la versión expuesta en la presentación de la práctica indicando que RAxML funciona mejor/es más preciso cuando el alineamiento es más grande, y que aplicar GTR con GAMMA es el modelo más preciso que se puede aplicar aunque el tiempo de computación sea mayor.

Observando los árboles también se puede apreciar una gran diferencia entre RAxML y FastTree, obteniendo unos resultados bastante más precisos con RAxML con un bootstrap casi perfecto. Por ello, el árbol elegido es el construido con RAxML con el modelo evolutivo GTR + GAMMA, el cual servirá para exponer los resultados obtenidos.
(Como poner el árbol entero es inviable, incluiré partes explicando algunas de las conclusiones obtenidas)



La primera cosa destacable es que la secuencia de referencia queda en una rama separada totalmente del resto de secuencias(hija izquierda de la raíz, resto de subárbol hijo derecho). Si bien cabría pensar que esto puede ser extraño, teniendo en cuenta las fechas de las secuencias, y que entre la de referencia y el resto hay aproximadamente un año, la similitud entre el resto es mucho mayor, de lo que se puede observar cómo el virus ha ido mutando desde un inicio considerablemente, y no es el mismo que empezó la pandemia.



Otro aspecto a comentar es la posición que ocupa la variante británica dentro del árbol. Cómo se puede observar no queda demasiado aislada del resto, de hecho esta por el medio del árbol y tiene varios “padres”, aunque si bien es cierto no se encuentra en el mismo nivel que ninguna otra, lo que puede sugerir que es una mutación notable del virus y que por eso se ha tenido tanto en cuenta.

Como conclusión, destacaría la rápida capacidad de mutación que tienen los virus, lo que se puede ver claramente en el árbol obtenido, en el que la secuencia de referencia está en otro subárbol a parte del resto de secuencias las cuales son más actuales.