Trabajo de Prácticas

Bioinformática

Fernando Peña Bes (NIA: 756012)

22 de abril de 2021

1. Introducción

El objetivo de el trabajo de prácticas de la asignatura es aplicar los conocimientos aprendidos en las prácticas anteriores para analizar un conjunto de secuencias de nucleótidos de SARS-CoV-2. Primero se realizará un alineamiento de todas las secuencias con la secuencia de referencia del virus, y a continuación se realizará un estudio de la variabilidad con el índice de conservación y se construirán varias árboles filogenéticos para explicar su historia evolutiva.

2. Desarrollo

2.1. Obtención de la información biológica

Se han obtenido todas las secuencias de alta calidad procedentes de Aragón y Cataluña que se encuentran en GISAID y a cada alumno le ha sido asignado un conjunto de unas 400 secuencias. A mi me han tocado secuencias de Cataluña recolectadas entre marzo y agosto de 2020, las más antiguas del conjunto de datos.

Hay un total de 391 secuencias con una longitud media de 29799.26 pares de bases. Las secuencias son de buena calidad, y no contienen muchas letras ambiguas, a continuación muestro el número medio de ocurrencias de cada una:

Código de nucleótido	Media de ocurrencias
A	8886.2404
C	5461.8031
T	9562.1586
G	5841.8491
N	46.9668
W	0.0153
S	0.0153
K	0.0332
Y	0.1535
R	0.0230
M	0.0026

Tabla 1: Media de ocurrencias de nucleótidos

El carácter N corresponde a cualquier nucleótido y los caracteres W, S K, Y, R y M, a dudas entre 2 nucleótidos [3].

Estas secuencias se van a comparar con el genoma de referencia de Wuhan, que tiene 29 903 nucleótidos, y con el genoma de referencia de la variante B.1.1.7 o cepa británica, que con una longitud de 29 764 nucleótidos (esta cepa presenta algunas deleciones).

Los primeros genomas de la variante B.1.1.7 se recolectaron en septiembre de 2020 en Reino Unido, y los primeros casos en España se reportaron en enero [4], por lo que es de esperar ninguna de las secuencias con las que se va a trabajar pertenezca a la variante.

2.2. Multialineamiento

Para poder analizar las secuencias en los siguientes apartados, se han realizado dos multialineamientos utilizando la herramienta MAFFT. El primero, que se usará en la Sección 2.3 es un alineamiento de todas las secuencias de Cataluña con la secuencia de referencia de Wuhan:

```
mafft --auto --thread 8  \
--addfragments gisaid_hcov-19_2021_03_23_catHC1.fasta RefSeqWuhan.fasta \
> alineamiento_2-3.fasta
```

El segundo igual pero añadiendo la secuencia de la variante británica al alineamiento:

```
mafft --auto --thread 8 \
--addfragments <(cat gisaid_hcov-19_2021_03_23_B117.fasta \
gisaid_hcov-19_2021_03_23_catHC1.fasta) RefSeqWuhan.fasta \
> alineamiento_2-4.fasta
```

Se ha usado la opción `--addfragments` para alinear todas las secuencias respecto a la de referencia. Este método, como se comentó en la Práctica 3 es muy rápido ya que al alinear respecto a una secuencia de referencia el coste se consigue reducir a $O(NL \log L)$, donde L es la longitud de la secuencia y N es el número de secuencias [1].

También se ha usado el flag `--auto`, que indica a MAFFT que elija el algoritmo de alineamiento que mejor se adapte a las secuencias de entrada. En ambos ha usado `FFT-NS-fragment`.

En el alineamiento final en ambos casos no aparecen gaps dentro de la secuencia de referencia, pero si que se añaden dos al final ya que algunas secuencias tenían dos adeninas más en la cola de poli(A).

2.3. Detección de mutaciones y errores de secuenciación

En este apartado se va a estudiar la variabilidad de nucleótidos en cada posición del alineamiento utilizando un índice de conservación basado en la entropía basada en [2].

Para calcular el índice de conservación de la posición i del alineamiento, primero se calcula la frecuencia con la que aparece cada nucleótido, v, en la columna i del multialineamiento de esta forma: $f_v(i) = n_v(i)/n(i)$. Una vez hecho esto, se calcula la medida de conservación según el método de entropía: $C(i) = -\sum_{v=1}^{4} f_v(i) \ln(f_v(i))$. La entropía es máxima cuando los cuatro nucleótidos aparecen con la misma frecuencia en la posición. Todos los caracteres distintos de A, G, C y T no se tienen en cuenta en el cálculo (los gaps tampoco). La implementación del cálculo del índice de conservación se encuentra en el script `indice_conservacion.py`.

Se ha realizado un estudio del índice de conservación para diferentes zonas del alineamiento, que ha consistido en calcular la media y obtener las cinco posiciones con mayor variabilidad. Los resultados se muestran a continuación:

```
Zona: Genoma completo, Posiciones: (sec ref: [1, 29903], alineamiento: [1, 29903])
Media de conservación: 0.00091
Posiciones con mayor entropía
  20268 0.68432
  26801 0.64088
  6286 0.62008
  445 0.61804
  28932 0.61724

Zona: Gen ORF1a, Posiciones: (sec ref: [266, 13483], alineamiento: [266, 13483])
Media de conservación: 0.00080
Posiciones con mayor entropía
  6286 0.62008
  445 0.61804
  5170 0.51625
  11132 0.51287
  11083 0.30141

Zona: Proteína Spike, Posiciones: (sec ref: [21563, 25384], alineamiento: [21563, 25384])
Media de conservación: 0.00074
Posiciones con mayor entropía
  22227 0.61450
  25314 0.17053
  23403 0.12794
  25049 0.09962
  23731 0.08958

Zona: Gen ORF3a, Posiciones: (sec ref: [25393, 26220], alineamiento: [25393, 26220])
Media de conservación: 0.00150
Posiciones con mayor entropía
  25563 0.13686
  25660 0.11877
  25591 0.08958
  25680 0.08958
  25886 0.05694

Zona: Gen N, Posiciones: (sec ref: [28274, 29533], alineamiento: [28274, 29533])
Media de conservación: 0.00275
Posiciones con mayor entropía
  28932 0.61724
  28883 0.49943
  28882 0.49565
  28881 0.47832
  29296 0.20890
```

La zona que presenta mayor media variabilidad es la correspondiente al gen N, aunque la del gen ORF1a es la que presenta más posición con variabilidad alta. La posición con mayor entropía de todo el alineamiento es la 20 268, que no se encuentra en ninguna de las zonas analizadas.

Para cada zona se ha realizado un gráfico de barras con el índice de conservación para cada posición (Figura 2.3). Se observa que índices de conservación para posiciones adyacentes suele ser muy distinto, aunque las zonas con más variabilidad tienden a agruparse

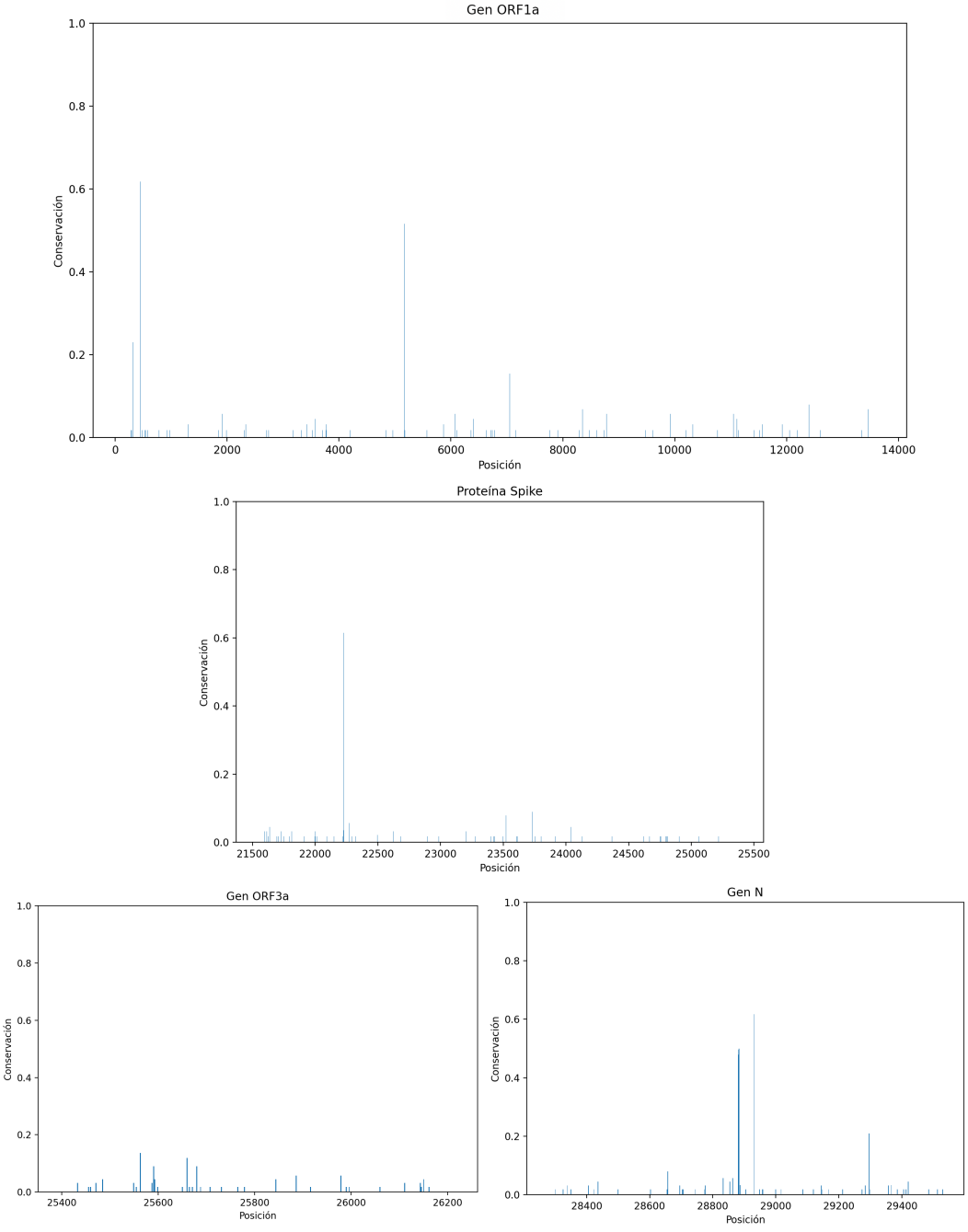

Figura 1: Medida de conservación de cada posición de las zonas analizadas

2.4. Construcción de una filogenia

Con el objetivo de explicar la historia evolutiva de las secuencias que se están analizando, se van a intentar construir un árbol filogenético. En esta práctica el método a utilizar va a ser el de máxima verosimilitud. Se basa en inferir la distribución de las secuencias en un árbol utilizando modelos evolutivos para que modelan la probabilidad de que ocurran mutaciones y el objetivo es encontrar el árbol más probable.

2.4.1. Herramientas

Las herramientas utilizadas han sido FastTree y RAxML, muy utilizadas para construir árboles filogenéticos, y MEGA X, con la que se trabajó en la práctica anterior.

En sistemas UNIX lo más sencillo para usar FastTree y RAxML es descargar su código y compilarlo. RAxML se puede descargar desde `https://github.com/stamatak/standard-RAxML` y compilar usando Make (se ha decidido compilar la versión con pthreads):

```
make -f Makefile.PTHREADS.gcc
```

En cuando a FastTree, se descargó el fichero `FastTree.c` de la web del programa, `http://www.microbesonline.org/fasttree/`, y se compiló de la forma que se indicaba:

```
gcc -O3 -finline-functions -funroll-loops -DUSE_DOUBLE -Wall \
    -o FastTree FastTree.c -lm
```

MEGA X está disponible en forma de instalador en `https://www.megasoftware.net`.

2.4.2. Pruebas resultados

Se ha trabajo con los modelos Jukes-Cantor y Generalized Time-Reversible (GTR). El primero es más sencillo que el segundo al asumir que todas las bases tienen la misma frecuencia y la misma tasa de sustituciones.

Para cada uno de las herramientas anteriores se ha intentado construir un árbol con cada uno de los modelos anteriores y se han recogido los resultados en la Tabla 2.

FastTree por defecto procesa alineamientos de proteínas, hay que usar el flag `-nt` para indicar que se trata de nucleótidos. Al usar este flag por defecto se usa el modelo Jukes-Cantor:

```
FastTree -nt alineamiento_2-4.fasta > fasttree_jc.nwk
```

Para seleccionar el modelo GTR basta con añadir el flag `-gtr`:

```
FastTree -nt alineamiento_2-4.fasta > fasttree_gtr.nwk
```

En RAxML el modelo GTR se puede elegir de la siguiente manera:

```
raxmlHPC-PTHREADS-AVX -s alineamiento_2-4.fasta \
    -n raxml_jc.nwk -m GTRCAT -p 222 -T 8
```

Usar Jukes-Cantor es una opción avanzada y hace falta indicarla añadiendo el flag `--JC69`, el modelo seleccionado con `-m` se ignora, pero es necesario ponerlo para que el programa no de errores al ejecutar:

```
raxmlHPC-PTHREADS-AVX -s alineamiento_2-4.fasta \
    -n raxml_gtr.nwk --JC69 -m GTRCAT -p 222 -T 8
```

Para comparar los árboles se usa la verosimilitud logarítmica del árbol resultante, *Log-likelihood score* en inglés.

Programa	Modelo	*Log-likelihood score*	Tiempo (s)
FastTree	Jukes-Cantor	-59447.870	142.68
FastTree	GTR	-56768.969	308.15
RAxML	Jukes-Cantor	-49900.702	102.41
RAxML	GTR	-48520.537	84.40
MEGA X	Jukes-Cantor	-49598.130	161.640
MEGA X	GTR	-48259.061	278.27

Tabla 2: Puntuaciones y tiempo de ejecución de los árboles

La mejor puntuación se ha obtenido con Mega, utilizando el modelo GTR. En la Sección 2.1 se ha comentado que las secuencias con las que se está trabajando son anteriores a la expansión de la variante B.1.1.7 del virus, además se ha comprobado con un script programado en la Práctica 3 que ninguna de las secuencias presenta las mutaciones típicas de esta variante, por tanto la secuencia de referencia de la variante británica tendría que aparecer separada del resto en el árbol filogenético. En el mejor árbol encontrado la secuencia de la variante británica está separada del resto desde la raíz como se muestra en la Figura 2. Además, la topología general del árbol y la distribución de los subárboles (Figura 3) es similar a la del que se encuentra en `nexstrain.org` para el periodo de tiempo en el que se recogieron las muestras (`https://nextstrain.org/ncov/europe?dmax=2020-8-31&dmin=2020-4-1`).

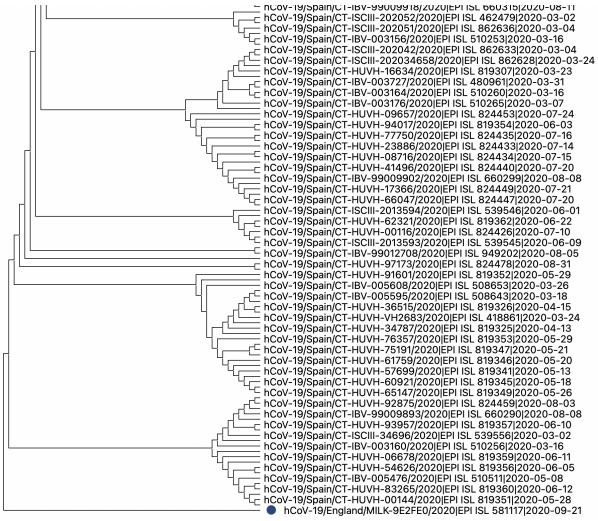

Figura 2: Variante británica en el árbol por MEGA con el modelo GTR.

Tanto FastTree como RAxML han tenido problemas al construir al topología del árbol a pesar de que todos los programas han usado un método similar de unión de vecinos para ello. Los resultados han árboles con subárboles muy pequeños y confusos, y en los que la secuencia de la variante británica aparecía muy cerca de las demás secuencias.

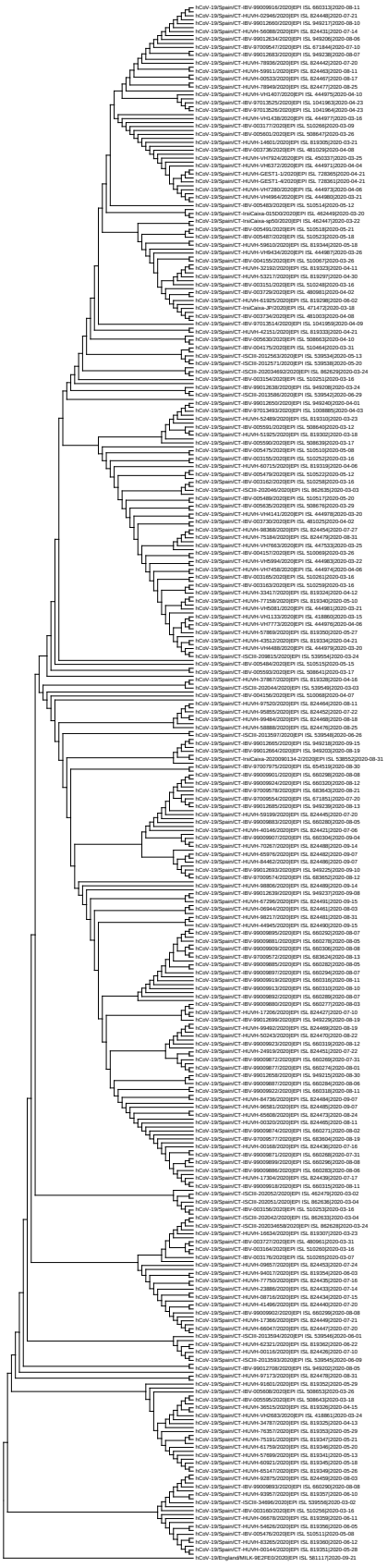

Figura 3: Árbol construido por MEGA con el modelo GTR.

Figura 4: Variante británica en el árbol por RAxML con el modelo GTR.

Referencias

[1] *MAFFT – Closely related viral genomes*. URL: https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html (visitado 16-03-2021).

[2] Jimin Pei y Nick V. Grishin. «AL2CO: calculation of positional conservation in a protein sequence alignment ». En: *Bioinformatics* 17.8 (ago. de 2001), págs. 700-712. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.8.700. eprint: https://academic.oup.com/bioinformatics/article-pdf/17/8/700/8201277/170700.pdf. URL: https://doi.org/10.1093/bioinformatics/17.8.700.

[3] *Wikipedia – Formato FASTA*. URL: https://es.wikipedia.org/wiki/Formato_FASTA (visitado 22-04-2020).

[4] *Wikipedia – Lineage B.1.1.7*. URL: https://en.wikipedia.org/wiki/Lineage_B.1.1.7 (visitado 21-04-2020).