Práctica 5:

Construcción de árboles con parsimonia y distancias, dos casos prácticos sencillos

Bioinformática

UNIZAR 2021

Grado de ingeniería informática

En esta cuarta sesión se comenzarán a utilizar herramientas de construcción de árboles filogenéticos, centrándonos en los métodos de máxima parsimonia y distancias.

Alejandro Gómez González

1. Obtención de información biolǵogica

En esta práctica los datos han sido directamente proporcionados junto al enunciado y se tratan de una serie de secuencias de nucleótidos de SARS-CoV-2 descargadas de GISAID[1].

En mi caso, se me han asignado asignadas una serie de secuencias provenientes de Cataluña.

2. Multialineamiento

Siguiendo las propias recomendaciones del portal web de MAFFT[2], se ha hecho uso del cálculo rápido para multi alineamientos completos de virus estrechamente relacionados[3] con las opciones por defecto. Este método es computacionalmente mucho menos costoso, y se debería de obtener igualmente un buen alineamiento.

Como referencia, en la prueba realizada para esta práctica, el multi alineamiento se ha realizado en menos de 20 segundos, teniendo en cuenta el tamaño de la muestra esta ejecución podría considerarse realmente eficiente.

A modo de evaluar la calidad del alineamiento, se puede tener en cuenta, por ejemplo, la diferencia entre la longitud final del alineamiento con respecto a la secuencia de mayor longitud del conjunto. Que en este caso la longitud es de 29903 en la secuencia original, y de 29910 en la secuencia de mayor longitud del conjunto, con una diferencia de tan solo 7 unidades.

También se ha calculado la puntuación del alineamiento de forma que se puntúa con 1 la coincidencia, -3 la diferencia, -5 el primer gap y -2 los siguientes gaps cuando haya varios seguidos para todas las secuencias y se hace la media de estas. El score obtenido en este caso es de 29140.2558 .

3. Detección de mutaciones y errores de secuenciación

El algoritmo de AL2CO[4] explicado brevemente en el guión ha sido implementado en python haciendo uso de la biblioteca de Bio para leer los datos en formato FASTA y NumPy para las operaciones con matrices necesarias.

---

[1] "GISAID - Initiative." https://www.gisaid.org/. Se consultó el 21 abr. 2021.
[2] "a multiple sequence alignment program - MAFFT." https://mafft.cbrc.jp/alignment/software/. Se consultó el 21 abr. 2021.
[3] "a multiple sequence alignment program - MAFFT." https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html. Se consultó el 13 abr. 2021.
[4] "AL2CO: calculation of positional conservation in a protein ... - PubMed." https://pubmed.ncbi.nlm.nih.gov/11524371/. Se consultó el 21 abr. 2021.

Junto a la memoria de la práctica se entrega un Jupyter Notebook (se puede abrir fácilmente en Google Colab) en el cual se encuentra todo el código desarrollado.

Para cada región se va a analizar su valor máximo encontrado, su valor mínimo, y su media. Además, se ha analizado toda la secuencia como punto de partida. Obteniendo los siguientes resultados:

```
WGS max: 0.6268997548740813
WGS min: -0.0
WGS mean: 0.003113392793408981

Gen ORF1A max: 0.621455593359433
Gen ORF1A min: -0.0
Gen ORF1A mean: 0.0030152218791885065

Protein Spike max: 0.6268997548740813
Protein Spike min: -0.0
Protein Spike mean: 0.004351872921077981

Gen ORF3A max: 0.21085605758676748
Gen ORF3A min: -0.0
Gen ORF3A mean: 0.005082016547965208

Gen N max: 0.6151547795735444
Gen N min: -0.0
Gen N mean: 0.0069974527714779106
```

Como se puede apreciar, el gen ORF1A no parece tener una variabilidad mayor o menor a el resto de la secuencia, sin embargo tanto en la proteína Spike, gen ORF3A y gen N se puede apreciar una variabilidad considerablemente mayor.

4. Construcción de una filogenia

Para la construcción de una filogenia se han probado las dos herramientas comentadas en la presetacion de la práctica: RAxML y FastTree

Primero, con RAxML se ha probado con el modelo de GTRCAT, que en el ordenador en el cual se realizó la práctica tardó 15 minutos. Su score fue de -54970.495768 .

Posteriormente, con FastTree, otra vez con el modelo GTR, se completó la tarea en 15 minutos también. Su score fue de -63215.488 .

Continuando con FastTree, esta vez con el modelo que usa por defecto, el JTT. Se completó la tarea en 3 minutos, su score fue de -66973.228 .

Después, se ha probado el modelo LG, completándose la tarea en 6 minutos, con un score de -66973.228

Finalmente, en FastTree también se ha probado el modelo WAG, completando la tarea en 5 minutos, con un score de -66973.228

Como se puede observar, no hay una opción que sea una clara vencedora, se debe en cada ocasión hacer un balance entre el tiempo de ejecución que queremos dedicar y la calidad de los resultados.

Se se desease un árbol lo más “correcto” posible, recomendaría hacer uso de RAxML con el modelo GTRCAT, sin embargo si se quiere obtener un resultado de forma rápida, usaría FastTree con su modelo por defecto JTT.