Práctica 4
Bioinformática

José Manuel Sánchez Aquilué 759267

18 de marzo de 2021

1. Introducción

La siguiente memoria recoge como se utilizaron programas de construcción de árboles filogenéticos en dos casos prácticos. Desde la obtención de las secuencias, hasta la visualización del árbol, pasando por el multialineamiento. Entre los distintos visualizadores, nos decantamos por MEGA, para consultar los resultados de Bootstrap, y por iTOL, que ofrece una forma más sencilla de visualizar y editar el árbol. En lo que respecta al formato, se ha trabajado en todo momento con árboles en Newick con sus visualizaciones en PNG, no ha sido necesario hacer ninguna conversión entre formatos de árboles.

El primer caso se trata de hacer un análisis filogenético del DNA del tigre de Tasmania y el de los felinos y los marsupiales. Se ilustrará con que grupo tiene más similitud en el DNA. Mientras que el segundo es una causa judicial de un médico que inyecto HIV a su novia. Nuestra labor consistirá en comparar las muestras de HIV de la damnificada, el paciente que atendía el procesado y el resto de pacientes del hospital, y así probar el delito del que se le acusa.

2. Caso 1: El tigre de Tasmania

2.1. Puesta a punto

Lo primero que debemos hacer es acudir a Genbank y descargar todas las secuencias de referencia de DNA mitocondrial de felinos y marsupiales. En principio trabajaremos con las 76 resultantes, siendo que no es un número demasiado grande. De esta manera, además, podremos llegar a conclusiones visualizando un único árbol de cada tipo, no hará falta consultar primero los felinos y luego los marsupiales. Si la cantidad de secuencias hubiese sido bastante más alta, no quedaría más remedio que desdeñar secuencias de cada tipo. La consulta a introducir sería la siguiente. Es importante establecer la configuración que se muestra en la figura 1.

```
1 (Felidae[Organism]) OR (Marsupialia[Organism])
```

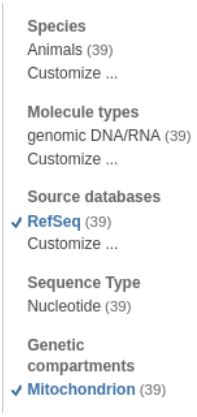

Figura 1: Búsqueda de DNA mitocondrial de referencia en Genbank.

El siguiente paso sería alinear las secuencias. Si bien MEGA permite alinearlas, no es recomendable, ya que los resultados arrojados al introducir ese alineamiento distan bastante del árbol filogenético correcto. Como alternativa, utilizaría mafft, que ya vimos en la práctica anterior que es bastante eficiente.

2.2. Generación de árboles filogenéticos

En este ejemplo se trabajará tanto con métodos de reconstrucción por distancia como basados en secuencia. Los basados en distancia son los dos vistos en clase, UPGMA y unión de vecinos. En cuanto al calculo de distancia, se probarán tres métodos distintos: número de diferencias, p-distancia y Jukes-Cantor. Por otro lado se utilizará el método máxima parsimonia con 50 iteraciones de bootstrap, lo mínimo que permite MEGA. Como se puede observar en los ficheros adjuntos a la práctica, el uso de bootstrap no condiciona la estructura del árbol en métodos de distancia. Ahora bien, nos facilita la fiabilidad estadística para comparar métodos. Del mismo modo, aumentar las iteraciones de bootstrap no altera la estructura.

En la entrega se han añadido las matrices de distancia de las secuencias del multialineamiento calculadas por MEGA. Dado que son matrices de 76x76 no se añadirán a la memoria, únicamente se mencionarán las formulas utilizadas.[1]

$$p - distancia : d = \frac{n_d}{n}$$

donde n_d es el número de diferencias y n el total de nucleótidos.

$$Jukes - Cantor : d = \frac{-3}{4} \ln\left(1 - \frac{3}{4}p\right)$$

donde p es la p-distancia.

En los árboles generados por UPGMA, se aprecia, independientemente del tipo de distancia utilizada, una gran bifurcación en el nodo raíz, que separa a los felinos de los marsupiales. A pesar del aspecto que tiene el tigre de Tasmania, y de su propio nombre, según su DNA, está más cerca desde un punto de vista evolutivo de los marsupiales que de los felinos. A modo de ejemplo, si prestamos atención al árbol generado (Figura 2), el tigre de Tasmania tiene bastante más en común con le numbat (Myrmecobius fasciatus) que con el tigre (Panthera tigris). Lo cual nos lleva a pensar que, al igual que el numbat, es un marsupial.

Si nos basamos en diferentes métodos de distancia podemos observar algunos cambios de orden en ciertos clados (Figura 3). Según la fiabilidad estadística obtenida al ejecutar el algoritmo con boostrap, dependiendo del tipo de distancia unos clados son más fiables que otros. No obstante, la zona del tigre de Tasmania tiene una fiabilidad del 100 % (Figura 4). Es decir, para superar el objetivo de esta práctica, por lo menos si se utiliza el algoritmo de UPGMA, da igual el tipo de distancia que se utilice.

Sería extraño que el programa se hubiese equivocado al clasificar al tigre de Tasmania. No obstante, vamos a probar otros métodos y a comparar la fiabilidad. En UPGMA hemos establecido como criterio $d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{r \in C_i, s \in C_j} d(r, s)$. Sin embargo, sería conveniente tener en cuenta el grupo que está muy lejos de los demás. La unión de vecinos consiste en minimizar $Q(i, j) = (r - 2)d(C_i, C_j) - u(C_i) - u(C_j)$, siendo $u(C_i) = \sum_j d(C_i, C_j)$.

La primera diferencia que apreciamos en este tipo de árboles, en comparación con los UPGMA, es la presencia de tres nodos hijos de la raíz (Figura 5). Eso se debe a que ha dividido a los felinos en dos subárboles, uno que contiene al genero panthera y otro con el resto. De todos modos, la separación de felinos y marsupiales continua vigente y, como era de esperar, el tigre de Tasmania figura como marsupial.

Aquí sí que se ve afectada la posición del tigre de Tasmania a la hora de cambiar la distancia con la que nos basamos. Además se observa una bajada en la fiabilidad estadística cuando la distancia utilizada es el número de diferencias. Con los otros tipos de distancia el resultado es similar al obtenido con UPGMA.

En ultimo lugar, vamos a generar el árbol filogenético siguiendo el método de máxima parsimonia. Este método parte de la hipótesis de que las mutaciones son poco frecuentes y construye un árbol que implique pocas mutaciones. La forma de este árbol es distinta a los demás, se puede contemplar una estructura como "escalonada" (Figura 7). Cuesta un poco más localizar la separación entre felinos y marsupiales, pero esta sigue apareciendo. En esta ocasión bootstrap muestra menos fiabilidad en el nodo del tigre de Tasmania, el resto de ramas adyacentes son altamente fiables.

Figura 2: Árbol generado con el algoritmo UPGMA basándose en el número de distancias.

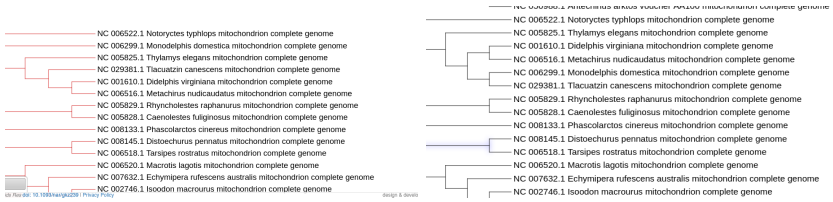

Figura 3: Mismo clado, a la izquierda calculado con número de diferencias y a la derecha con p-distancia.

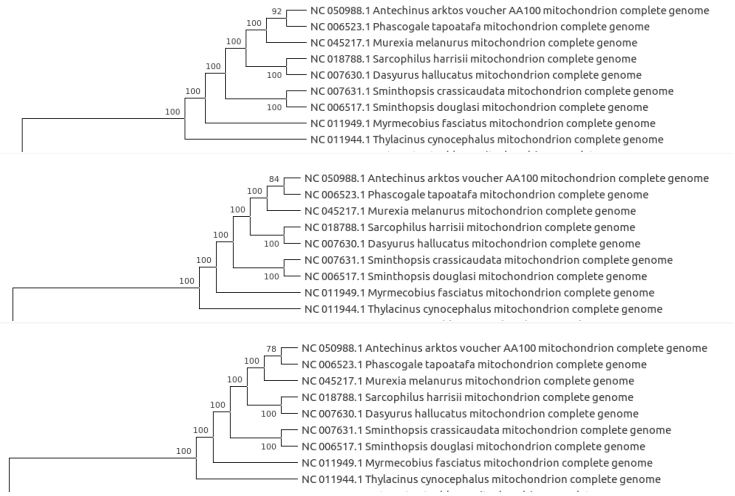

Figura 4: Clado donde se encuentra el tigre de Tasmania con la fiabilidad de cada nodo. Basandonos en los tres tipos de distancias n. diferencias, p-distancia y J-L (de abajo a arriba).

2.3. Conclusiones

La conclusión que obtenemos por unanimidad es que el tigre de Tasmania, pese a su aspecto, es un marsupial, y por ende un mamífero. En segundo lugar, todos los métodos son capaces de separar el DNA de los felinos y de los marsupiales en grupos diferentes y bien delimitados, lo cual nos indica que las especies de cada clase (entendida como grupo de especies, y no por el significado taxonómico de la palabra) tienen bastantes elementos en común en su DNA y difieren bastante del DNA de la otra clase.

Por otro lado, según el método utilizado, y la distancia, la estructura del árbol cambia. Por tanto, en algún otro tipo de problemas más complejos que requirieran definir las relaciones evolutivas perfectamente habría que dejarse guiar por distintas métricas para comparar y decantarse por el árbol más preciso.

Figura 5: Árbol generado por el algoritmo NJ.

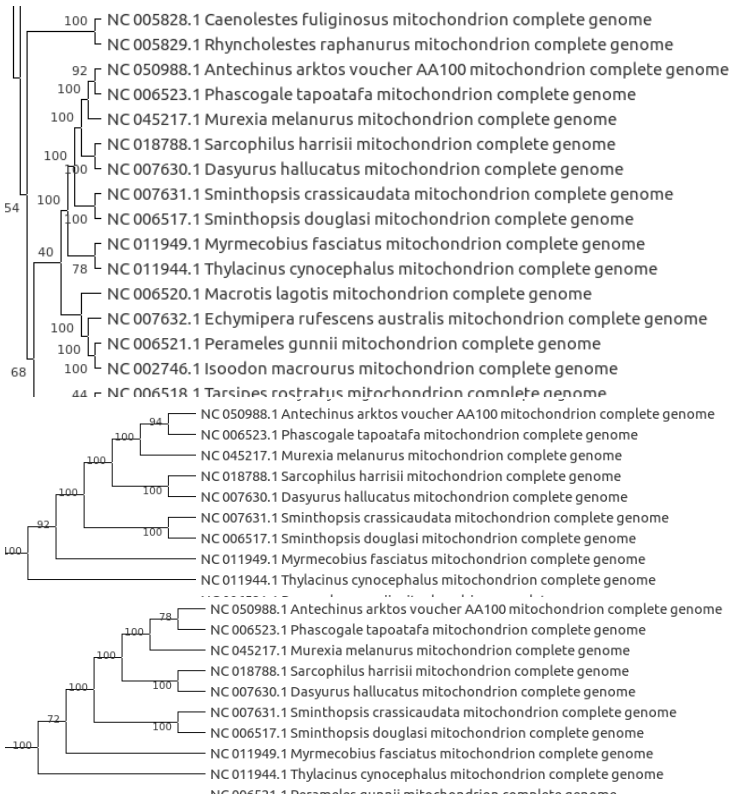

Figura 6: Clado donde se encuentra el tigre de Tasmania con la fiabilidad de cada nodo. Basandonos en los tres tipos de distancias n. diferencias, p-distancia y J-L (de abajo a arriba).

Figura 7: Árbol generado por el algoritmo de máxima parsimonia.

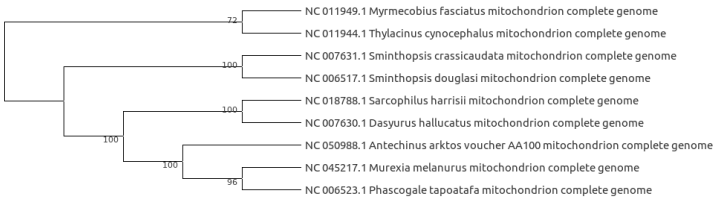

Figura 8: Fiabilidad estadística de los nodos cercanos al tigre de Tasmania del árbol obtenido por máxima parsimonia.

3. Caso 2: Resolviendo un caso criminal

3.1. Puesta a punto

Como no es exactamente el mismo alineamiento que en prácticas anteriores, sino que vamos a seleccionar 42 de ellas, lo mejor será volver a descargarlas y alinearlas. Volvemos a GenBank y a la consulta de la primer a práctica habría que añadirle como búsqueda de texto libre "(pol) gene", quedaría algo así:

```
1 ( AY156734 [ Accession ] : AY156907 [ Accession ] ) AND "(pol) gene"
```

El multialineamiento se realizará con mafft como en el caso anterior. Cabe recalcar, que me vi obligado a cambiar el nombre original de las secuencias, ya que este contenía pol entre paréntesis y se formaba una confusión con el formato Newick.

3.2. Generación de árboles filogenéticos.

Aunque en este ejemplo no haya una bifurcación desde el nodo raíz; en los árboles UPGMA, existe un nodo padre de todas las secuencias de la víctima y del paciente con el que trataba su novio (Figura 9). Independientemente de la distancia en la que nos basemos, el clado que contiene las secuencias mencionadas es idéntico, si bien es cierto que existe alguna ligera variación en el orden resto de secuencias. Ahora no tenemos una fiabilidad tan alta como la que había en el caso anterior, también es natural, ya que aquí las secuencias son todas del mismo virus y son bastante más parecidas. Vamos a ver si con los otros algoritmos damos con una clasificación distinta o con mayor fiabilidad.

Al aplicar la técnica de unión de vecinos, en función del tipo de distancia que utilicemos los resultados son variables. Todos comparten un denominador común, las secuencias de el paciente que atendía el novio de la víctima y las de la víctima tienen un padre en común. Nuevamente existe una separación perfecta entre las secuencias de el paciente y la víctima y el resto.

Para nuestra sorpresa, el método de máxima parsimonia no muestra esa división tan clara. Entre las secuencias de la víctima y del paciente, se encuentran

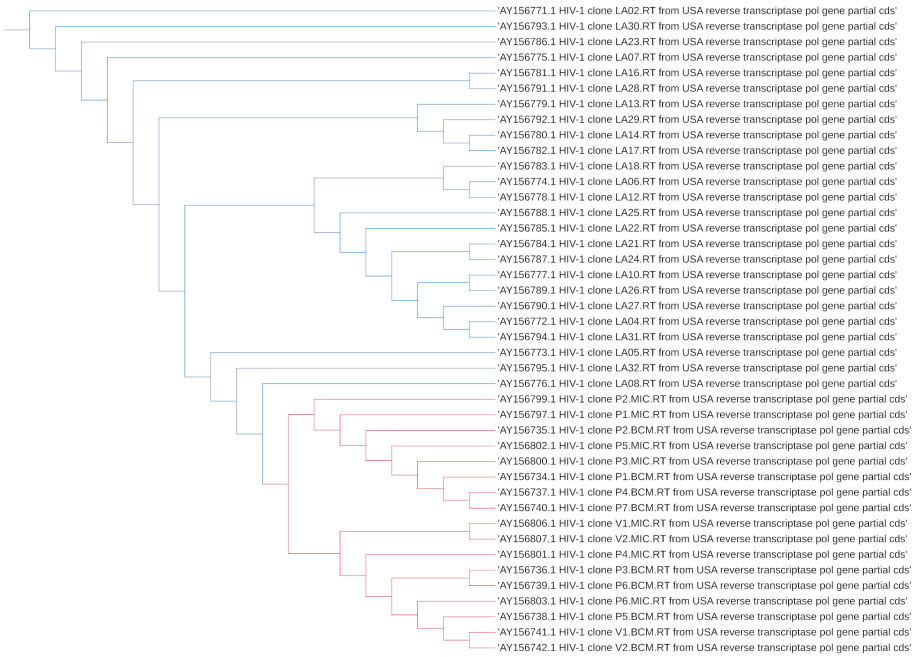

Figura 9: Árbol obtenido por el método de UPGMA

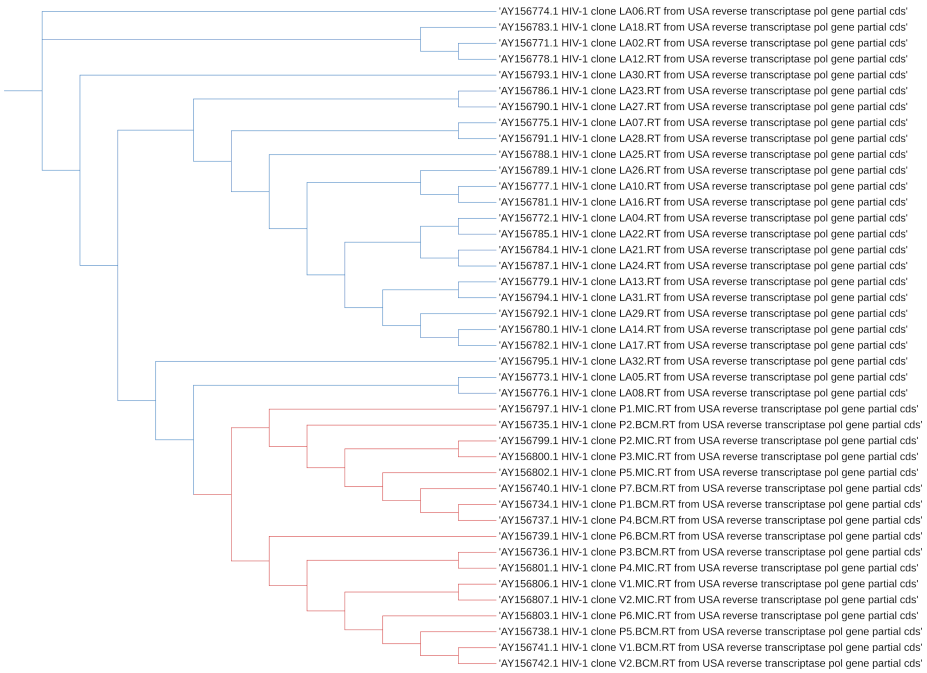

Figura 10: Árbol obtenido por unión de vecinos

secuencias del resto de pacientes(Figura 11). Con el fin de salir de dudas, consultamos la fiabilidad estadística. Efectivamente, con unos resultados tan bajos (Figura 11), no podemos aventurarnos a sacar ninguna clase de conclusión. Aumentando el número de iteraciones, lejos de conseguir más robustez, seguimos sin dar con una separación clara.

3.3. Conclusiones

Sin llegar a la concordancia entre métodos del caso anterior, podemos afirmar con bastante seguridad que el acusado es culpable de transmitirle el HIV a su novia. Los algoritmos basados en distancias muestran una similitud entre las secuencias de la víctima y las del paciente de su novio. Mientras que el cálculo basado en la máxima parsimonia no es del todo claro y arroja resultados confusos. En suma, si no nos llegamos a valer de diversos métodos de construcción de árboles y distintos tipos de distancias, existiría la posibilidad de que el acusado resultara inocente, al amparo del principio *in dubio pro reo*.

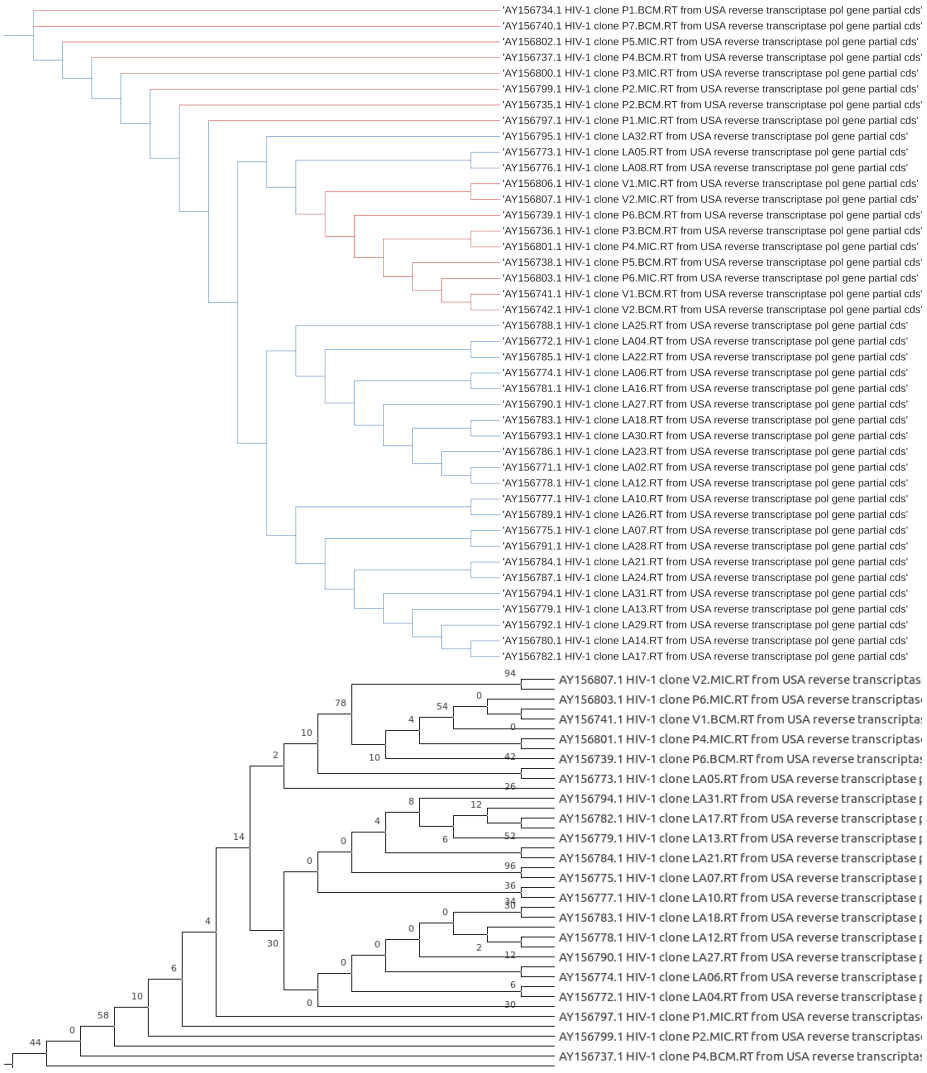

Figura 11: Árbol obtenido mediante máxima parsimonia junto su fiabilidad estadística

Referencias

[1] Koichiro Tamura Sudhir Kumar and Michael Li Glen Stecher, Glen Stecher. Documentación de mega. 2020.