Introducción a la Bioinformática

TP Especial - Parte 1

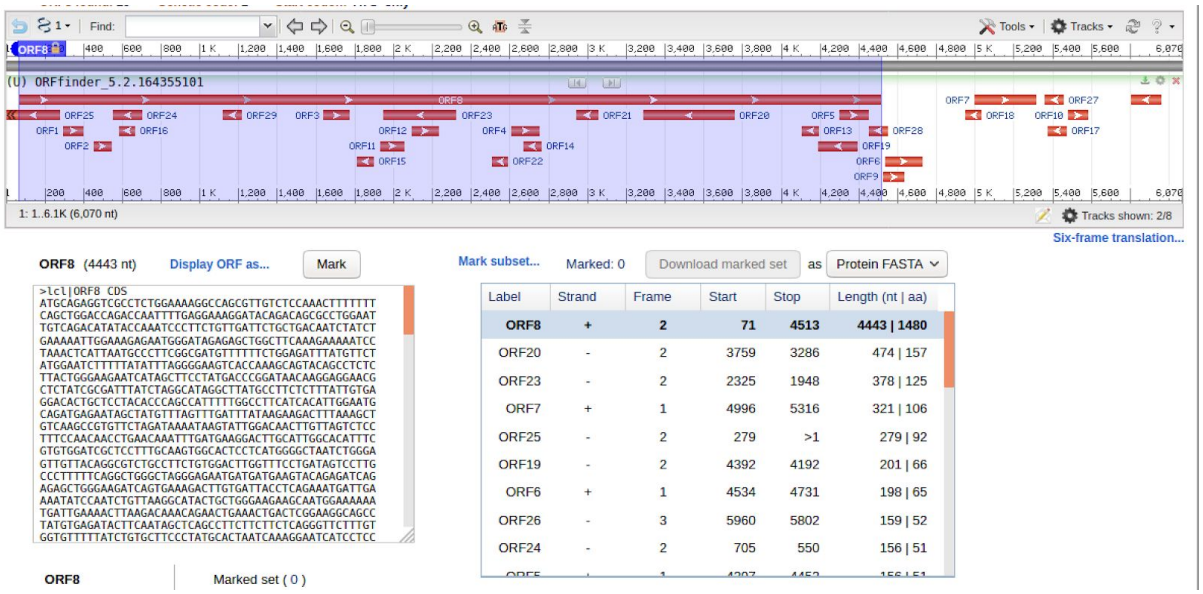
Francisco Pérez Sammartino
Mattéo Barbet
Martina Arco

TP Bioinformática

Para el trabajo 1 se decidió analizar la enfermedad genética Fibrosis Quística (Cystic Fibrosis).

Información sobre la enfermedad

La fibrosis quística (FQ) es un trastorno genético en la que hay acumulación de moco que obstruye algunos de los órganos del cuerpo, sobre todo los pulmones y el páncreas. Los síntomas pueden incluir un sabor salado de la piel, tos persistente, infecciones pulmonares frecuentes incluyendo neumonía o bronquitis, sibilancias o falta de aliento, crecimiento deficiente o pérdida de peso, frecuentes heces gordurosas y volumosas dificultad para evacuar e infertilidad masculina.
Con el tiempo, la acumulación de moco y las infecciones pueden conducir a daño pulmonar permanente, incluyendo la formación de tejido cicatricial (fibrosis) y quistes en los pulmones.
La FQ es causada por varios cambios (mutaciones) en el gen CFTR y se hereda de forma autosómica recesiva, es decir, que ambos genes de un par deben ser anormales para causar la enfermedad. Los tratamientos dependen de los síntomas, e incluyen terapia respiratoria, medicamentos inhalados, suplemento de enzimas pancreáticas, suplementos nutricionales y otros. Algunos medicamentos más recientes, los moduladores CFTR han sido aprobados para su uso en Estados Unidos. Los estudios de investigación en curso se centran en encontrar la cura para la enfermedad.
Aquí esta el link a la base de datos de OMIM: https://omim.org/entry/219700
Para ver el la secuencia del ARN mensajero del gen que está relacionado con la enfermedad, se utilizó la base de datos ncbi. En el siguiente link se encuentra la información del gen (se utilizó la información del RNA mensajero maduro):
https://www.ncbi.nlm.nih.gov/nuccore/1732746288
Esta secuencia de nucleótidos tendrá 6 marcos de lectura posible: 2 opciones por la dirección de lectura, y luego por cada dirección, 3 posibilidades para ver en qué posición se comienza a leer (1, 2 o 3). Si empiezo en la 4 es lo mismo que la primera ya que los codones son conjuntos de 3 nucleótidos.
Una vez tenida la secuencia, se usó ORF finder de ncbi para poder ver cual era el marco de lectura correcto (a pesar de que la decisión del marco de lectura podría ser incorporada dentro del programa). El resultado, basándose en intentar obtener la cadena más larga conociendo los codones de inicio y fin, fue el siguiente:



Es decir que la longitud de la secuencia es de 4443 nucleótidos (o 1480 aminoácidos), que comienza en la posición 71 y termina en la 4513. El codón de inicio en este caso fue ‘ATG’ y el de finalización ‘TAG’.

Herramientas a utilizar:

- El proyecto será realizado en el lenguaje Python.
- Se utilizara el paquete BioPython (https://biopython.org/) para realizar el manejo de las secuencias, el blast y las demás operaciones relacionadas con la bioinformática especificamente.

Ejercicios

Los ejercicios se realizaron en el lenguaje python utilizando la librería de biopython para el manejo de la información biológica.

Ejercicio 1

En este ejercicio se recibe como entrada el archivo descargado de la base de datos que contiene la secuencia de nucleótidos del ARN mensajero maduro. Se utilizó el siguiente arn mensajero: “Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA” [https://www.ncbi.nlm.nih.gov/nuccore/1732746288]
El objetivo del programa es traducir a la secuencia de aminoácidos, para ello se debe determinar el marco de lectura correcto (ORF) y el comienzo y final de la codificación del gen.
Consideramos que la traducción correcta es aquella más extensa, es decir, con más nucleótidos, y que comience por Metionina (AUG) y finalice en alguno de los 3 codones de stop(UAG/UGA/UAA).

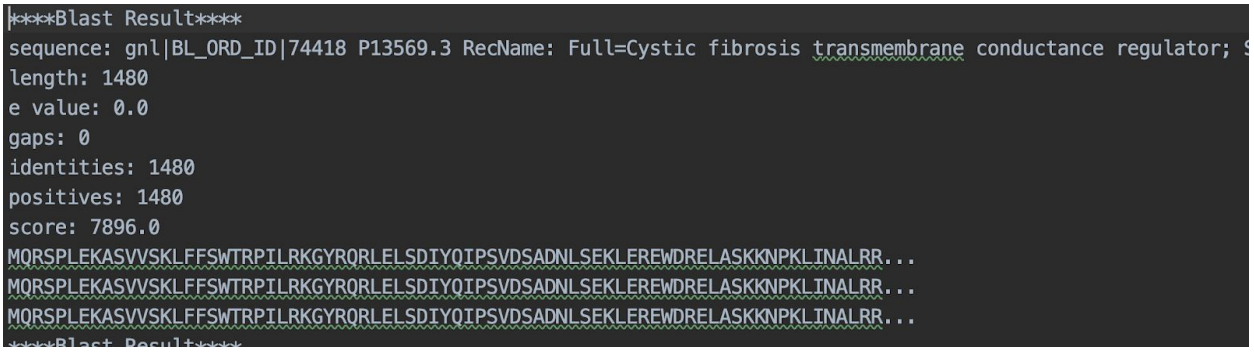
En otras palabras, iteramos por cada uno de los marcos de lectura (3 en el sentido original y 3 en el reverso), y vamos guardando la secuencia más larga encontrada por el momento. Al finalizar, obtendremos la más larga.
Una vez obtenida la secuencia más larga, se la traduce a aminoácidos. (en el algoritmo se traduce la secuencia cuando se la almacena como la más extensa actualmente en las iteraciones del algoritmo).
El resultado obtenido al analizar y traducir la secuencia previamente descrita, fue el mismo que al realizar la traducción en la web de NCBI, confirmando la correctitud del algoritmo.

Ejercicio 2

El objetivo de este ejercicio es el de realizar el blast sobre la secuencia de aminoácidos obtenida como resultado del ejercicio anterior. En el programa están habilitadas ambas formas de procesamiento, local y online. Para el procesamiento local se debe contar con la base de datos Swissprot descargada y formateada previamente.
Tanto de forma local como online se utilizó blastp para realizar el alineamiento.
En los resultados se pueden observar los siguientes valores estadísticos:

- **Length**: longitud de la secuencia encontrada
- **E Value**: está relacionado con la cantidad de hits que uno puede encontrar en la base de datos “de casualidad”. Si es cercano a 0, implica que el resultado o el match es más significativo.
- **Gaps**: Es un espacio que se introduce en el alineamiento para compensar inserciones o borrados en una secuencia relativos a otra.
- **Identities**: indica la cantidad de posiciones en que coinciden los aminoácidos
- **Positives**: indica la cantidad de posiciones donde el score de sustitución es positivo. (si son iguales es positivo)
- **Score**: Se calcula como la suma de los gaps (tienen valores negativos según la longitud del gap) y los valores de sustitución (positivos o negativos) que se ven en una tabla de tipo BLOSUM.

Para el caso del trabajo, al realizar el alineamiento se obtuvo como primer resultado, el mismo gen del cual se partió, que era lo esperable. Los resultados fueron los siguientes:



La longitud de la secuencia a alinear era de 1480, y la encontrada coincide. El valor de identities coincide con la longitud ya que todas las posiciones son idénticas. Esto nos indica que ambas secuencias son iguales.
El valor de *e value* es 0 indicandonos la relevancia del match; esclarece que el match no tiene nada de casualidad. La posibilidad de encontrar un match con ese score de forma aleatoria es 0.
Luego de este primer HIT, se pueden ver otros resultados, muchos con score muy cercano al primero que pertenecen a otras especies. En orden de aparición en los resultados las especies son las siguientes:

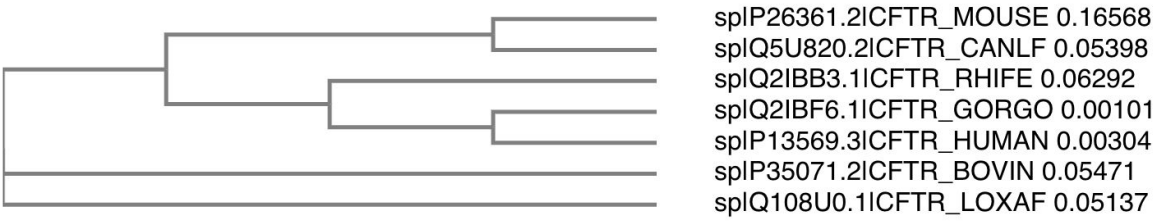
Score	Nombre Especie	Nombre conocido
7865	Gorilla gorilla gorilla	Gorila
7862	Pan troglodytes	Chimpancé Común
7822	Pongo abelii	Orangután de Sumatra
7818	Nomascus leucogenys	Gibón de mejillas blancas
7763	Macaca mulatta	Mono Rhesus
7762	Papio anubis	Papión oliva
7758	Macaca nemestrina	Macaco Cola de cerdo
7754	Chlorocebus aethiops	Cercopiteco Verde
...	...	..
6983	Canis Lupus Familiaris	Perro
6943	Loxodonta africana	Elefante Africano
...	...	
6873	Bos taurus	Vaca
6865	Rhinolophus ferrumequinum	Murciélago grande de Herradura
...	...	...
6064	Mus Musculus	Ratón Casero
5936	Xenopus laevis	Rana de uñas africana
5910	Rattus norvegicus	Rata parda

Ejercicio 3

En este ejercicio se espera recibir como entrada un archivo MSA, con un conjunto de secuencias, y con ello realizar un alineamiento múltiple que permite encontrar similitudes y diferencias entre las diferentes secuencias.
Las secuencias de entrada deben estar conformadas por la secuencia de consulta original, es decir la del gen relacionado con la fibrosis múltiple en humanos, y con la secuenciación del mismo gen pero en otros organismos.
Es interesante que a partir del alineamiento podremos obtener el árbol filogenético que nos mostrará la relación evolutiva del gen entre las diferentes especies.
Los organismos elegidos para el alineamiento de sus secuencias de proteínas son:

- **Humano**, Homo Sapiens (CFTR_HUMAN):
  - https://www.ncbi.nlm.nih.gov/protein/P13569.3
- **Gorilla**, Gorilla Gorilla (CFTR_GORGO):
  - https://www.ncbi.nlm.nih.gov/protein/Q2IBF6.1
- **Perro**, Canis Lupus Familiaris (CFTR_CANLF):
  - https://www.ncbi.nlm.nih.gov/protein/Q5U820.2
- **Elefante Africano**, Loxodonta Africana (CFTR_LOXAF):
  - https://www.ncbi.nlm.nih.gov/protein/Q108U0.1
- **Vaca**, Bos Taurus (CFTR_BOVIN):
  - https://www.ncbi.nlm.nih.gov/protein/P35071.2
- **Murciélago**, Rhinolophus ferrumequinum (CFTR_RHIFE):
  - https://www.ncbi.nlm.nih.gov/protein/Q2IBB3.1
- **Ratón Casero**, Mus Musculus (CFTR_MOUSE):t
  - https://www.ncbi.nlm.nih.gov/protein/P26361.2

Luego de realizar el MSA se pudo obtener el árbol Filogenético que se muestra a continuación. Para obtener el árbol se utilizó el aplicativo web que se encuentra en el siguiente link: http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=clustalw2_phylogeny&sequence=clustalw2-I20141008-205527-0685-78599923-es
Allí se tomó como input la salida del ejercicio 3, es decir el alineamiento entre las 7 secuencias y como resultado se obtuvo el árbol.



Se puede ver que el Elefante Africano, y la Vaca son aquellos que tiene la codificación más antigua de la proteína, es decir que ha sufrido menos mutaciones. El humano es aquel cuya proteína ha sufrido más cambios junto con el Gorila. La diferenciación con el Gorila se hace muy cercana a las hojas del árbol, de ahí que el gorila sea el segundo resultado al realizar blast con la proteína humana.
Un resultado a remarcar es que el murciélago, que tenía un score peor que la Vaca, el Perro y el Elefante en el blast, está cercano en el árbol, es decir que tiene un ancestro en común con el humano más reciente.
Los más alejados evolutivamente son la vaca y el elefante africano, ya que el ancestro en común se encuentra más lejano que con las demás especies.
Se puede notar que en la raíz del árbol salen 3 ramas, esto se denomina politomía. Esto sucede cuando no hay información suficiente para determinar el orden de las ramificaciones. Recordemos que solo existen ramificaciones en 2 ramas y no en 3.

Bibliografía

Información sobre la enfermedad: https://rarediseases.info.nih.gov/