Introducción a la Bioinformática

(16.50)

Trabajo Práctico Especial
Parte I

Julián Palacci - 57157
Agustín Lavarello - 57046
Santiago Swinnen - 57258

Introducción

En el presente informe se presentará y desarrollará el trabajo práctico especial realizado en el marco de la materia Introducción a la Bioinformática del Instituto Tecnológico de Buenos Aires en el primer cuatrimestre de 2019.
El trabajo, desarrollado en el lenguaje Python con el soporte de la librería BioPython, consta de tres partes.
La enfermedad elegida para el presente trabajo fue cáncer de mama, la cual está muchas veces asociado al gen **BRCA2**

Ejercicio 1

Para realizar el ejercicio 1, dada una o varias secuencias de nucleótidos, se busco para cada una los marcos de lectura abiertos (ORF) tanto en la secuencia original, como en la secuencia complementaria reversa. Para cada marco de lectura se realizó la traducción a la secuencia de aminoácidos correspondiente.
Para determinar cuál era el marco de lectura correcto, se buscó el marco de lectura más largo que comenzara en una M (metionina).
Para este ejercicio se utilizaron expresiones regulares para identificar codones de inicio (ATG).
Se utilizó el mRNA asociado al gen BRCA2.

Ejercicio 2

En este ejercicio, se realizan las consultas blastp de la secuencia de proteínas. Los valores estadísticos que arroja el blast son:

- E-value: Es el número de coincidencias que uno puede esperar encontrar de casualidad cuando busca en una base de datos. Cuanto más cercano a cero sea el E-value, más significativo es el match, aunque las secuencias cortas suelen tener e-values altos debido a que tienen más probabilidad de ocurrir en una base de datos de casualidad
- Score: Se calcula como la suma de los scores de sustitución y gaps. Los scores de sustitución están determinados por la matriz de sustitución (BLOSUM)
- Positives: Es el número de aminoácidos que son iguales entre las secuencias analizadas o que tienen propiedades químicas similares
- Gaps: Es número de espacios que se le agregan a la secuencia para compensar inserciones o borrados en una secuencia relativa a la otra

El resultado del blast muestra que la traducción de la proteína que supusimos fue la correcta, pues se observa que Breast cancer type 2 susceptibility protein matchea en forma completa, es decir su e-value es de 0, no tiene gaps, y las identities y los positives coinciden con el length de la secuencia input. Esto indica que la traducción a la proteína que se hizo en el ejercicio 1, es correcta.
Además, se puede observar que hay otras proteínas similares en otras especies que se asocian a la misma enfermedad como *Felis catus*, *Mus musculus, Rattus norvegicus*

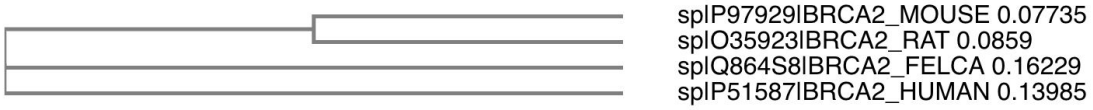
Ejercicio 3

El programa generado por el archivo *ex3.py* espera dos argumentos. En primer lugar un archivo multifasta como input, y en segundo lugar un archivo de output. Para realizar el alineamiento múltiple, se utiliza Clustal Omega.

Una vez obtenidos los alineamientos, se escriben en el archivo de output y el programa finaliza.

El MSA permite alinear tres o más secuencias de proteínas, ADN o ARN. Se asume que las secuencias que se quieren alinear tienen una relación evolutiva y que comparten descendientes de un mismo ancestro. Del resultado del MSA, se pueden inferir secuencias homólogas y realizar un análisis filogenético para poder determinar los ancestros comunes de las especies involucradas. En base a este análisis es posible luego construir un árbol filogenético.

Realizando el alineamiento a través de ClustalW pudimos obtener el árbol filogenético de las especies



Con lo que se observa que *Mus musculus* y *Rattus norvegicus* comparten un ancestro en común y a su vez ese ancestro comparte un ancestro en común con *Felis catus* y *Homo sapiens*