Fecha límite de entrega del TP completo (Partes 1 y 2): 8 de Junio de 2020

Introducción a la Bioinformática
Trabajo Práctico – Parte 2

Ejercicio 4 – BLAST OUTPUT. Escribir un script para analizar (*parsear*) un reporte de salida de blast que identifique los hits que en su descripción aparezca un Pattern determinado que le damos como parámetro de entrada. El pattern puede ser una palabra. Nota para punto extra: Si quieren pueden parsear cuál es el ACCESSION del hit seleccionado (donde hay una coincidencia del Pattern) y con el módulo Bio::DB::GenBank obtener la secuencia completa del hit en formato FASTA y escribirla un archivo, es decir, levantar la secuencia original de los hits seleccionados.

- **Input: Reporte Blast (blast.out del ej. 2) y un Pattern (por ej. "Mus Musculus").**
- **Output: Lista de los hits que coincidan con el pattern (por ej. solo los hits de Ratones).**

Deben entregar el script Ex4.pm y su input file con una breve descripción.

Ejercicio 5 - EMBOSS. Instalar EMBOSS. Escribir un script que llame a algún programa EMBOSS para que realice algún análisis sobre la una secuencia de nucleótidos fasta (del Ej. 1). Por ejemplo, que calcule los ORF y obtenga las secuencias de proteínas posibles. Luego bájense los motivos de las bases de datos PROSITE (archivo prosite.dat) y por medio del llamado a otro programa EMBOSS realizar el análisis de dominios de las secuencias de aminoácidos obtenidas y escribir los resultados en un archivo de salida.

- **Input : Archivo de secuencias Fasta (ej. Xxxxx.fas con una o más secuencias de aa.**
- **Output: Archivo de resultados del dominios encontrados en las secuencias de aa.**

Ejercicio 6. Trabajo con Bases de Datos Biológicas (útil para la presentación de la investigación Ej. 8).

a) A partir del gen o proteína de interés para ustedes dar su link a NCBI-Gene como una entrada de Entrez, por ej.: http://www.ncbi.nlm.nih.gov/gene/3630
Expliquen brevemente lo que hace la proteína y por qué la eligieron.

b) ¿Cuántos genes / proteínas homólogas se conocen en otros organismos? Utilicen la información que está en la base de datos de HomoloGene y en la bases de datos Ensembl . Describan los resultados en ambas bases de datos, y en qué se diferencian. Mencionen sobre qué tan común creen son estos genes o proteínas y a qué grupos taxonómicos pertenecen (sólo en las bacterias, en los vertebrados, etc.)

c) ¿Cuántos transcriptos y cuántas formas alternativas de *splicing* son conocidos para este gen / proteína? ¿Cuáles de estos *splicing* alternativos se expresan? ¿Tienen funciones alternativas? Buscar evidencia de esto en las base de datos de NCBI y en los transcriptos de Ensembl ¿Cómo el número de *splicings* alternativos diferente entre las dos bases de datos y cuál piensan que es más precisa y por qué?

d) ¿Con cuántas otras proteínas interactúa el producto génico de su gen? ¿Existe un patrón o relación entre las interacciones? Mencione las interacciones interesantes o inusuales. Usted encontrará las interacciones de su gene/proteína tanto en la base de datos NCBI Gene como en la base de datos UniProt . Compare las dos tablas entre sí. ¿Hay proteínas que interactúan únicas para cada tabla?

e) Expliquen brevemente de qué componente celular forma parte su proteína (pista: se puede estudiar la información de Gene Ontology - GO), ¿A qué procesos biológicos pertenece (pista idem)? y ¿En qué función molecular trabaja esta proteína? Los términos ontológicos de genes los pueden encontrar tanto en NCBI Gene y en la base de datos UniProt como haciendo una búsqueda en AmiGO.

f) Discutan brevemente en qué estructura o vías metabólicas específicas (*pathways*) estaría participando su gen / proteína? (Reactome, KEGG son algunas bases de datos de pathways).

g) Entrar en la base de datos de variantes genéticas dbSNP e intentar interpretar o encontrar info sobre alguna variante (reference SNP - rsXXXX) asociada con la patología investigada en su gen de interés. ¿Qué variante es? ¿Hay información sobre la frecuencia que tiene esta variante en la población? ¿Qué grupo étnico parece ser el más afectado?

NOTA: Para hacer este ejercicio les pueden servir algunas otras bases de datos como:
http://www.genecards.org
http://www.ncbi.nlm.nih.gov/clinvar/ (para obtener información clínica del gen y sus variantes)
https://ghr.nlm.nih.gov

Ejercicio 7. A partir de las secuencias de recibidas de SARS-CoV-2 de pacientes infectados por COVID-19 en la Argentina, realicen una mini investigación bioinformática para intentar obtener información de interés biológico y/o epidemiológico para este conjunto de cepas.

El ejercicio es libre, pueden desarrollar cualquier tarea de investigación bioinformática que se sientan cómodos para ejercitar los conocimientos adquiridos en la materia.

Ejercicio 8. Armar una presentación donde expliquen la enfermedad que investigaron, lo que hicieron y los resultados que fueron obteniendo en los ejercicios del TP.

Los integrantes de cada grupo tendrán un máximo de 10 minutos para exponer como realizaron el trabajo práctico y comentar sobre sus investigaciones (no tanto sobre el código implementado). La correcta exposición del trabajo realizado por los miembros del grupo también entra en la evaluación.