Introducción a la Bioinformática - Trabajo Práctico Nro. 1

Fecha de entrega 23 de Junio 2014
Grupos de hasta 4 estudiantes. Egresados deben formar grupo entre ellos.

El presente trabajo práctico tiene por objetivo adquirir las primeras habilidades en el campo de la Bioinformática. Se incluyen cuatro ejercicios donde deberán desarrollar pequeños scripts para resolver problemas específicos. Los mismos pueden ser desarrollados utilizando cualquiera de los lenguajes de programación bioinformática de código abierto como BioPerl, BioJava y BioRuby, que son ampliamente utilizados en la investigación bioinformática y de biología computacional. Las herramientas computacionales escritas en estos lenguajes proporcionan múltiples funcionalidades para crear soluciones personalizadas y realizar análisis de datos biológicos. Un quinto ejercicio estar relacionado con la comprensión de la información en bases de datos de biología molecular.

Para comenzar el trabajo práctico deben entrar en la base de datos *Online Mendelian Inheritance in Man* (OMIM) donde encontrarán el catálogo online genes humanos asociados a trastornos genéticos más importante de la actualidad. Decidan sobre que enfermedad quieren investigar y luego seleccionen uno más genes asociados a esta para comenzar con el ejercicio 1. Este mismo gen o genes seleccionados deben utilizarse en el ejercicio 5.

Cada grupo tendrá 15 minutos para exponer como realizó el trabajo práctico y comentar sobre su investigación. Por favor preparen una presentación. La correcta exposición del trabajo realizado por los miembros del grupo también entra en la evaluación. Los 3 minutos finales tendrá lugar el concurso de muestra de habilidades GeneBoy entre los grupos. Un ejemplo de muestra de habilidades: Cargo la secuencia del gen de interés, obtengo los ORF y sus secuencias aminoácidos posibles, las comparo con Blastp a La db de Uniprot y veo alineamiento a la secuencia de la proteína real. Suerte a todos los concursantes!

- **Ejercicio 1 – PROCESAMIENTO DE SECUENCIAS**. Escribir un script que lea una o más secuencias (de nucleótidos) de un archivo que contenga la información en formato GenBank de su gen (o genes) de interés, las transcriba a sus secuencias de amino ácidos posibles y escriba el resultado en un archivo formato FASTA. Ustedes deben generarse su archivo GenBank de secuencias input, por ejemplo realizando una consulta por el gen INS que está asociado a la Diabetes en la base de datos de NCBI-Gene o mismo en la base de datos de Nucleótidos y obtener uno o más resultados en formato GenBank en un archivo de texto.

  - Input: Archivo de secuencias Genbank (ej. Xxxxx.gbk con una o más secuencias).
  - Output: Archivo de secuencias Fasta (ej. Xxxxx.fas con una o más secuencias de aminoácidos).

Deben entregar el script Ex1.pm (si lo hacen con BioPerl, sino será otra extensión) y el input file que utilicen con una breve descripción de lo que hicieron y como se debe ejecutar para probarlo.

- **Ejercicio 2 - BLAST.** Escribir un script que realice un BLAST de una o varias secuencias (si son varias se realiza un Blast por cada secuencia input) y escriba el resultado (blast output) en un archivo. Nota: Pueden ejecutar BLAST de manera remota o bien localmente (si hacen ambos tienen más puntos!), para esto deben instalarse BLAST localmente del FTP del NCBI, luego bajarse la base de datos ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/swissprot.gz y descomprimirla en un dir por ej. ncbi-blast-2.2.27+/data/, luego usar el comando ncbi-blast-2.2.27+/bin/makeblastdb sobre el archivo swissprot (el original ya está en formato FASTA) para darle formato de BLAST DB.

  - Input: Secuencia Fasta (por ej. Xxxxx.fas con una o más secuencias de aa obtenidas del Ej.1).

  - Output: Repote Blast (por ej. blast.out, si deciden hacer múltiples pueden generar un único o varios archivos).

Deben entregar el script Ex2.pm y su input file con una breve descripción de lo que hicieron y como se debe ejecutar para probarlo.

- **Ejercicio 3 – 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 modulo 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."Arabidopsis").

  - Output: Lista de los hits que coincidan con el pattern (por ej. solo los hits de Arabidopsis).

Deben entregar el script Ex3.pm y su input file con una breve descripción.

- **Ejercicio 4 - EMBOSS.** Instalar EMBOSS. Escribir un script que llame a algún programa EMBOSS para que a partir de una secuencia de nucleótidos fasta (del Ej. 1) 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 aminoácidos).

  - Output: Archivo de resultados del dominios encontrados en las secuencias de aa.

- **Ejercicio 5. Trabajo con Bases de Datos Biológicas.**

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? ¿Cúales 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).

- Entregar un documento de texto con las respuestas.