Fecha límite de entrega: última clase de Noviembre de 2015

Introducción a la Bioinformática
Instalación de Linux - BioPerl - Blast Suite

Para poder desarrollar el trabajo práctico deberán tener instalados el sistema operativo Linux, el lenguaje de programación Perl con las librerías BioPerl y los programas de Blast.

Linux
Pueden realizar una partición del disco o bien instalar una máquina virtual.
Las distribuciones Linux que suelen utilizarse en bioinformática son Debian o Ubuntu. Existen distribuciones como Bio-Linux cuyos paquetes pueden ser instalados sobre Debian o Ubuntu, o la máquina virtual de DNALinux, ambas ya tienen Perl preinstalado, EMBOSS y otras herramientas que les serán de utilidad ya incluidas.

Perl
Los programas Perl son llamados scripts y tienen extensión *.pl (o bien *.cgi si son aplicaciones web). Perl es una lenguaje interpretado o de scripting (aunque también hay compiladores) cuya estructura deriva del C y toma cosas de la programación shell.
Instalación y documentación en: http://www.perl.org

BioPerl
BioPerl es proyecto comunitario open source de módulos Perl integrados para trabajar con secuencias y anotaciones, acceder a bases de datos remotas, parsear el output de programas como BLAST, FASTA, etc. Es prácticamente esencial para todo bioinformático. BioPerl-core tiene los módulos principales, BioPerl-run es una colección de módulos que facilitan la ejecución de programas locales como EMBOSS suite. Instalar BioPerl desde http://www.bioperl.org

Blast local (BLAST+ is a new suite of BLAST tools that utilizes the NCBI)
Existen distintas maneras de correr Blast de manera local:
Desde línea de comandos, previa instalación desde:
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
> blastp -d swissprot -i demo.fasta -o myblast.report (con standalone Blast)
> blastall –p blastp -d swissprot -i demo.fasta -o myblast.report (con Blast+)

BLAST HELP:
http://www.ncbi.nlm.nih.gov/books/NBK52637/

Desde BioPerl:
Utilizando el objeto Bio::Tools::Run::StandAloneBlast.pm (con Blast instalado local)
Utilizando el objeto Bio::Tools::Run::RemoteBlast.pm (usando el Blast del NCBI)

Introducción a la Bioinformática
Trabajo Práctico (final)

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, aunque se sugiere la utilización de BioPerl para facilitar la resolución de los ejercicios. Las herramientas computacionales escritas en estos lenguajes proporcionan múltiples funcionalidades para crear soluciones personalizadas y realizar análisis de datos biológicos. Un quinto ejercicio esta 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. En grupo decidan sobre que enfermedad quieren investigar y luego a partir de la información en OMIM seleccionen uno más genes asociados a esta patología para comenzar con el ejercicio 1. Este mismo gen o genes seleccionados deben utilizarse en el ejercicio 5.

Cada grupo tendrá 10 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.

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 un mRNA de su gen (o genes) de interés, las traduzca a sus secuencias de amino ácidos posibles (tener en cuenta los Reading Frames) y escriba los resultados en un archivo en formato FASTA. Ustedes deben generarse su archivo GenBank de secuencias input, por ejemplo realizando una consulta de los mRNA del gen INS (que está asociado a la Diabetes) en la base de datos de NCBI-Gene y obtener uno o más resultados en formato GenBank en un archivo de texto. Si no desean seguir trabajando con las seis secuencias de aa posibles, pueden utilizar alguna función o programa que les permita saber cual el es marco de lectura correcto y seguir con esa secuencia.
NOTA: Ver aclaración de este ejercicio al final del documento.

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

- Output: Archivo de secuencias Fasta de cada ORF (ej. Xxxxx.fas con una o más secuencias de aminoácidos (aa)).

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. Dependiendo de la versión de Blast suite que tengan instalado puede que en vez de makeblastdb deban utilizar el comando formatdb.

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

- Output: Reporte 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, con una interpretación de los resultados del Blast, y mencionar como se debe ejecutar para probarlo.

Ejercicio 2 (opcional) – Multiple Sequence Alignment (MSA). Descargarse las secuencias fasta de los 10 mejores resultados Blast y realizar un alineamiento múltiple con su secuencia de consulta más estas 10. Intenten realizar una interpretación del resultado del alineamiento múltiple.

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 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."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 aa.

- 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).

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.genemania.org

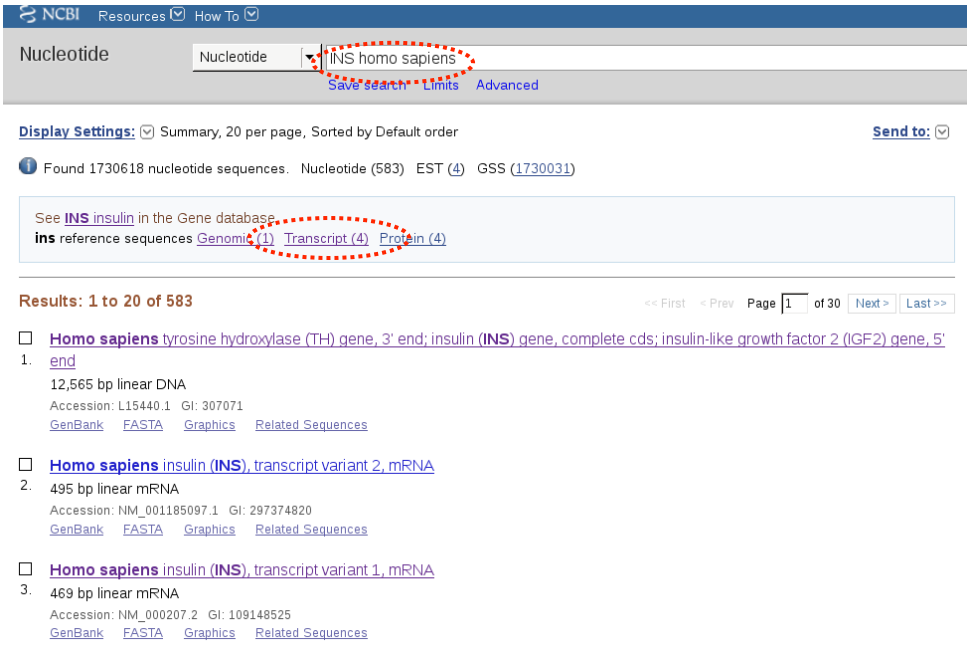
- Entregar un documento de texto con las respuestas.

Aclaración para el Ejercicio 1:

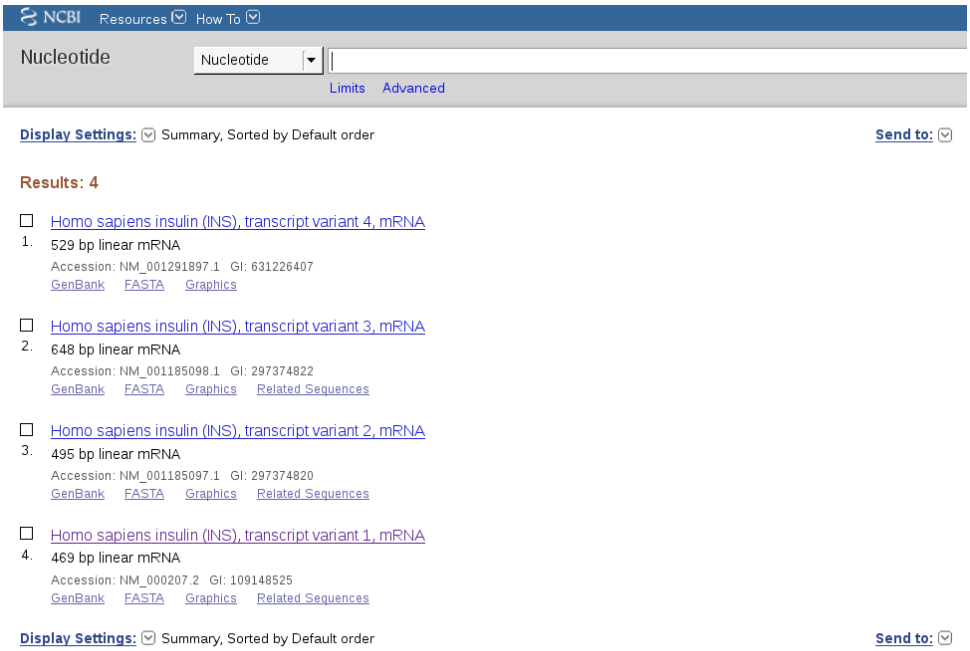
Para bajar una secuencia de nuestro gen elegido que funcione para en el Ejercicio 1 y para los demás, deberán bajarse alguno de los RNA mensajeros maduros (transcripto) de su gen. Es decir, una secuencia de mRNA que ya haya sido procesada y no tenga intrones, esta es la secuencia que deben bajarse en formato Genbank y hacer la traducción a su secuencia de aminoácidos.

Por ejemplo, para el gen de la insulina humano (INS):

1. Hago la búsqueda en la base de datos de nucleótidos:

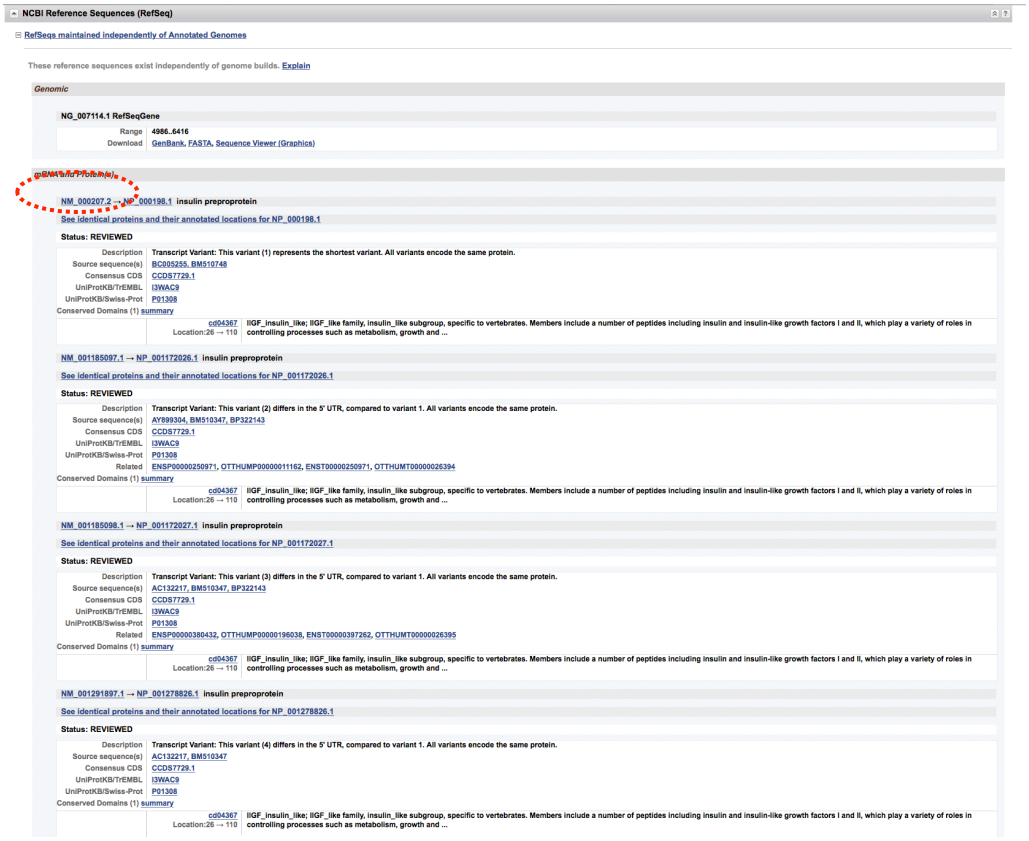


2. Selecciono los resultados de Transcript (4):

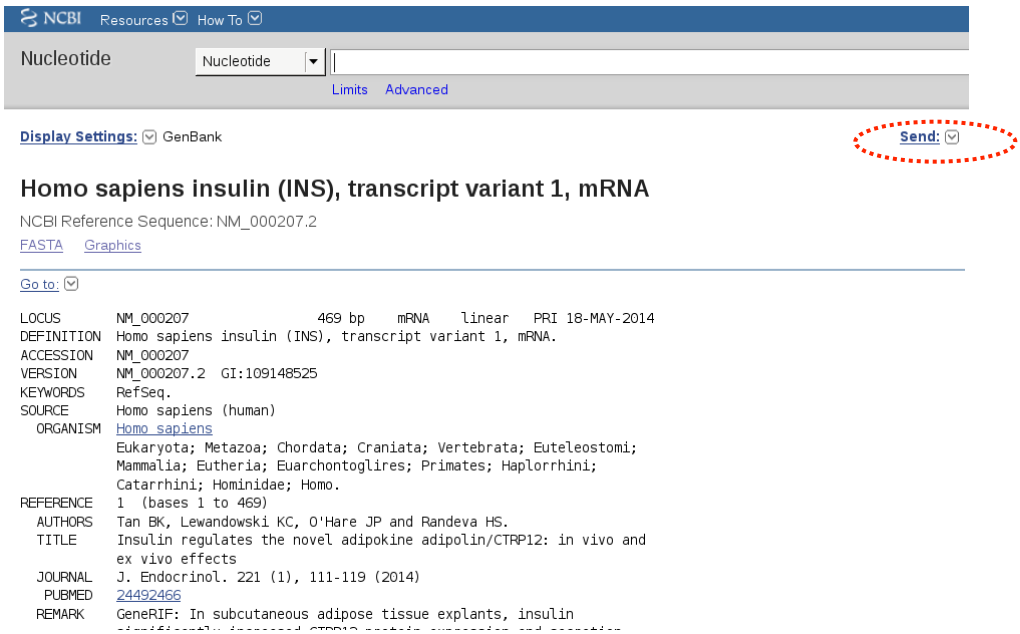


1. y 2. de otra forma (mejor)... como lo hicimos en clase

O bien una búsqueda en la base de datos de genes e ir a las secuencias de mRNA (NMxxxx)



3. Selecciono uno de los transcriptos del gen en formato GenBank (en lo posible el variant 1):



4. ORF (Open Reading Frame)

Una vez que tienen la secuencia bajada tengan en cuenta que ustedes desconocen cuál es el marco de lectura correcto de los 6 posibles. Por lo tanto deberán calcular los marcos de lectura posibles y evaluar todos ellos en el ejercicio 2, y así darse cuenta cuál de los 6 es el real. Existen funciones en BioPerl para hacer esto.