BLAST
EJERCICIOS

Ejercicio 1

Haz una búsqueda *blastp* en el NCBI usando la siguiente secuencia de búsqueda:

```
PNLHGLFGRKTG
```

Por defecto, los parámetros se van a reajustar para búsquedas de secuencias cortas. Inspecciona el resumen de la búsqueda. ¿Cuál es el umbral para el E-valor? ¿Cuál es el tamaño de palabra? ¿Cuál es la matriz de puntuación? ¿Cómo son estos valores con respecto a los valores por defecto, y en qué afectan los cambios?

Ejercicio 2

Las búsquedas por proteínas generalmente dan más información que las de nucleótidos. Haz una búsqueda *blastp* para RBP4 (NP_006735), restringiendo la búsqueda a "Arthropods" (insectos).

Luego, haz la misma búsqueda para nucleótidos con *blastn* (NM_006744; selecciona sólo los nucleótidos correspondientes a la región codificante)

¿Qué búsqueda nos da más información?

¿Cuántas coincidencias tienen un E-valor menor que 1.0?

Ejercicio 3

Descarga *blast+* del NCBI. Esta es la herramienta para uso local del BLAST del NCBI. La descarga e instalación tarda apenas unos segundos, y es una herramienta muy útil para hacer búsquedas BLAST muy grandes, que podrían sobrecargar el servidor web del NCBI.

blast+ funciona sólo desde la línea de comandos.

1. Crea un fichero *hbb.fa* que contenga la hemoglobina beta humana en formato fasta:

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

2. Invoca BLAST desde la línea de comandos para realizar una búsqueda en la base de datos no redundante (nr):

```
blastp -query hbb.fa –remote -db nr -out hbb.txt
```

   - `blastp` es el programa (igualmente tenemos blastn, blastx, etc.)
   - `–query`, `–remote`, `-db` y `–out` indican distintos parámetros, seguidos de los valores que toman (si necesitan valor)
     - `–query` indica el fichero de entrada (hbb.fa)
     - `–remote` indica que la base de datos a utilizar va a buscarla en los servidores de NCBI (si no usamos este argumento buscará la base de datos en el directorio donde está la query)
     - `–db` indica la base de datos (nr en nuestro caso)
     - `–out` indica el fichero de salida (lo llamaremos *hbb.txt*)

Inspecciona la salida que resulta (*hbb.txt*).

¿Cuál es el E-valor del mejor alineamiento?

¿Cuál es el espacio de búsqueda? (aparece al final del documento, en el resumen de los parámetros)?

La opción `–searchsp` sirve para limitar el espacio de búsqueda. Ejecuta de nuevo con `–searchsp 40000000` (40 millones) y guarda en otro fichero la salida (*hbb2.txt*).

Para descargar blast+

1. Descargamos blast+ de ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
2. Tenemos su manual aquí: http://www.ncbi.nlm.nih.gov/books/NBK1762/
3. Descomprimimos a la carpeta que queramos
4. En dicha carpeta introducimos el fichero hbb.txt, y ejecutamos la orden vista (o cualquier otra)

Posibles problemas:

1) No encuentra el programa

```
melchior:bin rodri$ blastp -query ../../blast/hbb.fa -remote -db nr -out test_query.fa
-bash: blastp: command not found
```

Estamos en la carpeta correcta, pero no encuentra el programa, dando un error tipo “command not found”.

En Linux (y si usas Mac, por debajo también tienes un sistema tipo Unix) necesitamos indicar la carpeta en la que estamos para ejecutar algo. En este caso, la carpeta actual (se expresa con su ruta completa, o con "."), con lo cual:

```
./blastp -query blast/hbb.fa -remote -db nr -out test_query.fa
```

debería funcionar.

2) No encuentra el fichero con las secuencias de consulta

```
melchior:bin rodri$ ./blastp -query hbb.fa -remote -db nr -out test_query.fa
Command line argument error: Argument "query". File is not accessible:  `blast/hbb.txt'
```

Asegúrate de que la ruta al fichero es correcta, seguramente *hbb.fa* no esté en la misma carpeta que el resto de blast, o tenga otro nombre

Opcional: podemos configurar nuestro perfil de usuario para no tener que ejecutar BLAST desde su carpeta de instalación:

- Windows: pinchar con el botón derecho en el icono de "Mi PC" o "Equipo", seleccionar "Propiedades", y luego la pestaña "Avanzada", botón "Variables de entorno…". Allí, en variables del sistema, buscamos PATH, la seleccionamos y le damos a "Editar…". En el campo "valor de la variable", sin borrar lo que haya, añadimos un ; al final, y después la dirección de la carpeta donde hemos instalado blast. Salimos dando a "Aceptar" a todo y reiniciamos el ordenador. Ahora desde el terminal podremos ejecutar blast desde cualquier ubicación.
- Linux/Mac: en la carpeta del usuario (podemos volver a ella con el comando "cd"), podemos listar los ficheros ocultos (empiezan por un punto), y tendremos alguno de nombre profile o similar (en Mac es .profile, en algunos Linux .bash_profile, etc.). Debemos abrir ese fichero con un editor de texto, y al final del mismo añadir la línea *export PATH=$PATH:/ruta/completa/al/directorio/de/instalación/de/blast/bin*. Es muy importante que apuntemos a la carpeta "bin" dentro del directorio de instalación. Guardamos el fichero y reiniciamos el terminal para poder ejecutar blast desde cualquier ubicación.

3) No encuentra la base de datos de búsqueda:

```
melchior:blast rodri$ blastp -query hbb.fa -db nr -out test_query.fa
BLAST Database error: No alias or index file found for protein database [nr]
in search path
[/Users/rodri/Documents/docencia/bioinformatica/programas/blast::]
```

En principio, usaremos las bases de datos del NCBI, por lo que tenemos que especificar el parámetro `-remote`. Si queremos usar bases de datos locales, debemos bajarnos la base de datos que queramos del NCBI antes, y especificar la ruta completa a su ubicación en el parámetro `-db`

4) Errores con el formato de entrada

Puede que veamos errores raros tipo éste:

```
melchior:blast rodri$ blastp -query hbb.docx -db nr -out test_query.fa -remote
Error: NCBI C++ Exception:
    "/am/ncbiapdata/release/blast/src/2.2.25/IntelMAC-universal/c++/GCC401-ReleaseMT-
-IntelMAC-universal/../src/objtools/readers/fasta.cpp", line 592: Error:
ncbi::objects::CFastaReader::CheckDataLine() - CFastaReader: Input not marked as defline
or comment, but contains too many special characters to be plausible data (m_Pos = 1)
```

Generalmente esto ocurre si nuestro texto de entrada contiene caracteres raros. NO editéis los ficheros de secuencias con programas de edición complicados tipo Word u OpenOffice, pues añaden caracteres especiales sobre el formato, que no necesitamos. Utilizad en su lugar NotePad (Windows), gedit (Linux) o TextEdit (Mac), o similares.

Ejercicio 4

Blast+ es una buena opción para hacer trabajos en segundo plano o "*batch*" (es decir, dejar muchas búsquedas BLAST ejecutándose, sin tener los resultados inmediatamente).

Vamos a hacer una búsqueda con blast+ de tres proteínas: beta globina humana, odorant-binding protein bovina, y citocromo b para el parásito de la malaria *Plasmodium falciparum*:

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|129022|sp|P07435|OBP_BOVIN    Odorant-binding    protein    (OBP)
(Olfactory mucosa pyrazine-binding protein)
AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQTTELTELFVKLNVEDEDLEKFWKLTEDKG
IDKKNVVNFLENEDHPHPE
>gi|11466247|ref|NP_059668.1| cytochrome b [Plasmodium falciparum]
MNFYSINLVKAHLINYPCPLNINFLWNYGFLLGIIFFIQIITGVFLASRYTPDVSYAYYSIQHILRELWS
GWCFRYMHATGASLVFLLTYLHILRGLNYSYMYLPLSWISGLILFMIFIVTAFVGYVLPWGQMSYWGATV
ITNLLSSIPVAVIWICGGYTVSDPTIKRFFVLHFILPFIGLCIVFIHIFFLHLHGSTNPLGYDTALKIPF
YPNLLSLDVKGFNNVIILFLIQSLFGIIPLSHPDNAIVVNTYVTPSQIVPEWYFLPFYAMLKTVPSKPAG
LVIVLLSLQLLFLLAEQRSLTTIIQFKMIFGARDYSVPIIWFMCAFYALLWIGCQLPQDIFILYGRLFIV
LFFCSGLFVLVHYRRTHYDYSSQANI
```

Guarda las tres secuencias en un archivo 3proteins.fasta y realiza una búsqueda blastp:

```
blastp -query 3proteins.fasta –remote -db refseq_protein -out 3proteins_out.txt
```

El fichero de salida 3proteins_out.txt contendrá los resultados para tres búsquedas blastp, una por cada secuencia fasta encontrada en el fichero de entrada.

Ejercicio 5

La familia de genes humanos más grande se dice que es la del receptor olfativo. Haz una búsqueda BLAST (en NCBI) por esta familia.

Pista: Una estrategia es ir a Entrez Gene y buscar por "olfactory receptor AND Homo sapiens[Organism]". Esto retorna unos 3000 registros, pero no nos dice que estén relacionados, así que selecciona uno de ellos (alguno que sea muy representativo) y haz una búsqueda BLASTP, restringida a humano, para ver cuáles de los 3000 forman familia significativamente.

Ejercicio 6

En el ejercicio 5, qué pasa si cogemos una matriz más adecuada para secuencias muy distintas (BLOSUM45)?

Ejercicio 7

¿La proteína de HIV-1 *pol* está mas relacionada con la proteína *pol* de HIV-2 o con la proteína *pol* del virus de inmunodeficiencia de los simios (SIV)?

Combina Entrez Gene para buscar los identificadores con blastp para determinar la similitud.

Ejercicio 8

El conocido como "hombre de hielo" ("iceman") es un espécimen que vivió hace 5300 años y cuyo cuerpo se recuperó de los Alpes italianos. Cierto material fúngico se recuperó de sus ropas, y fue secuenciado. ¿A qué moderno hongo se parece más el ADN de dicho material fúngico?

Ejercicio 9

Al hacer una búsqueda con BLAST, una de las coincidencias tiene un E-valor de 1e-4. ¿Qué significa? ¿De qué parámetros depende?

Ejercicio 10

Vamos a crear una proteína artificial, consistente en la proteína RBP4 seguida del dominio C2 de la proteína kinasa Ca (PKCA) en humano. Luego, hacemos una búsqueda PSI-BLAST. ¿Detectamos los múltiples dominios? ¿Hay alguna proteína que tenga tanto el dominio de la lipocalina como el C2?

Ejercicio 11

¿Existen hemoglobinas en los hongos? Haz una búsqueda PSI-BLAST usando la beta globina humana (NP_000509), restringiendo la salida a hongos, en la base de datos nr.

¿Cuántas coincidencias con un E-valor menor que 0.005 aparecen? ¿Y en una segunda iteración de PSI-BLAST? ¿Y con BLAST normal?

¿Qué tamaño medio tienen las proteínas fúngicas que contienen un dominio de globina?

Ejercicio 12

Vamos a usar PHI-BLAST. Busquemos la RBP4 humana (NP_006735) en bacterias (sólo refseq). Utilizamos el patrón PHI `GXW[YF]X[VILMAFY]A[RKH]`

Repite la búsqueda con el patrón `GXW[YF][EA][IVLM]`

Compara ambas búsquedas ¿Cómo varían los resultados? ¿Cuáles son los E-valores?

¿Los resultados mejoran respecto a un blastp normal?