CURSO: CC471

Laboratorio Dirigido 01.

1. Introducción

PRACTICA : BASES DE DATOS BIOLOGICAS.

Objetivo general:
-Conocer y utilizar las bases de datos primarias que se utilizan en Bioinformática.

Objetivos particulares:

- -Aplicar técnicas eficientes para la búsqueda de referencias bibliográficas en PubMed.
- -Conocer y aplicar diferentes técnicas para localizar y descargar archivos de secuencias nucleotídicas y de proteínas en las bases de datos del NCBI.
- -Aplicar técnicas para explorar detalladamente las regiones de genes y genomas registrados en el NCBI.

Nota: Genere un reporte (documento)con sus respuestas a las preguntas formuladas en las secciones "Desrrollo" y "Ejercicios de Evaluación" <CC471-Nobre-Apellido-Lab01.doc> subalo en formato comprimido al sitio web de recursos del curso: (mega.nz) en la carpeta "tareas".

INTRODUCCION.
Para que los datos de biología molecular (secuencias y estructuras biológicas) colectados en los últimos años tengan un impacto real en el desarrollo de las ciencias relacionadas con la biología, estos deben organizarse y depurarse para integrar bases de datos. Una base de datos es un sistema informático que almacena datos y proporciona herramientas para la consulta y organización de los mismos.

En la actualidad las bases de datos desempeñan un papel muy importante para la investigación biológica ya que suelen constituir las primeras fuentes de información que se consultan para estudiar un tema determinado. Las bases de datos empleadas en bioinformática pueden dividirse en bases de datos de información documental y en bases de datos relacionadas con moléculas biológicas.

Existen diversas bases de datos de información documental, algunas de ellas de acceso libre entre las que destaca PubMed. La base de datos PubMed es un servicio de la National Library of Medicine (NLM) que incluye más de 16 millones de citas bibliográficas provenientes de MEDLINE y de otras publicaciones científicas del área biomédica recopiladas desde los años cincuentas.

PubMed incluye vínculos con artículos completos y con otras fuentes de información relacionadas.

En cuanto a las bases de datos de moléculas biológicas, en general estas se dividen en primarias y secundarias. A principios de los años ochenta, la información sobre secuencias colectadas en la literatura comenzó a ser muy abundante y por tal motivo resultó conveniente almacenar dichos datos en bases de datos. Estos proyectos iniciales de almacenamiento de datos de secuencias y estructuras de moléculas biológicas constituyeron las denominadas bases de datos primarias entre las que destacan para secuencias nucleotídicas el GenBank mantenido actualmente por el National

Center for Biotechnology Information (NCBI), la base de datos del European Molecular Biology Laboratory (EMBL) y la del DNA Database of Japan (DDBJ).

Para proteínas las bases de datos primarias son las del Protein Information Resources (PIR), la del Swiss Center of Bioinformatics (SWISS-PROT) así como las secciones de secuencias derivadas de la traducción de las secuencias nucleotídicas codificantes del NCBI (GenPept) y del EMBL (TrEMBL).

Para datos de estructuras tridimensionales de proteínas y ácidos nucleicos hay que destacar la base de datos del Protein Data Bank (PDB) y la Nucleic Acid Database (NDB) respectivamente. Estas bases de datos son en la actualidad los reservorios principales de los datos de secuencias y estructures de moléculas biológicas provenientes de todos los organismos estudiados en el mundo.

Por otra parte existe una gran variedad de bases de datos secundarias o especializadas las cuales colectan información relacionada con un número reducido de organismos en particular o con algún proceso biológico determinado. El número de estas bases de datos secundarias aumenta rápidamente cada año y es prácticamente imposible dar una descripción detallada de todas las existentes.

La revista Nucleic Acids Research publica cada año un artículo en línea en el cual se proporciona una lista de las bases de datos primarias y secundarias más importantes para el área biológica. La actualización para el año 2014 contiene una lista de 1,552 bases de datos y el número de estas se incrementa notablemente cada año.

En esta práctica se estudiarán algunas de las técnicas más importantes para llevar a cabo la consulta de bases de datos primarias para la obtención de información sobre publicaciones, así como para la descarga de secuencias biológicas y estructuras.

RECURSOS INFORMÁTICOS
http://www.ncbi.nlm.nih.gov/ (Nacional Center for Biotechnology Information – NCBI)
http://www.ebi.ac.uk/ (European Bioinformatics Institute- EBI- EMBL)
http://www.ddbj.nig.ac.jp/ (DNA Databank of Japan- DDBJ)
http://www-nbrf.georgetown.edu (Protein Information Resource- PIR)
http://us.expasy.org/sprot/ (Swiss-Prot)
http://www.rcsb.org/pdb/home/home.do (Protein Data Bank- PDB)
http://ndbserver.rutgers.edu/ (Nucleic Acid Database – NDB)

DESARROLLO

I.Consulta de referencias bibliográficas en PubMed.
PubMed es una de las bases de datos del NCBI que puede ser consultada mediante el sistema de administración de bases de datos relacionales conocido como Entrez. Para ingresar a PubMed, se debe entrar a la página principal del NCBI. Dicha página muestra en la parte superior una lista desplegable en la cual se tienen accesos a algunas de las bases de datos más importantes administradas por el NCBI (Una lista más completa de las bases de datos del NCBI puede obtenerse desplegando la lista).

Presionando PubMed en la lista desplegable de bases de datos se ingresa a la página principal de esta base de datos. Para ilustrar como se pueden realizar búsquedas eficientes utilizando este sistema de consulta se utilizará como ejemplo la obtención de referencias bibliográficas relacionadas con la secuencia de la hemoglobina.

1. En la línea para búsquedas teclear la frase “hemoglobin sequence” y presionar el botón “Search”. Observar el número de referencias obtenidas y la forma como son clasificadas por el sistema. Examine algunos de los títulos de los trabajos encontrados ¿Considera apropiado el número de referencias encontrado? ¿Considera que las referencias encontradas son relevantes para el problema en estudio?

2. Realizar la búsqueda anterior pero utilizando operadores lógicos ó booleanos (AND, OR, NOT) para refinar las búsquedas. Los operadores deben ser escritos siempre con mayúsculas para que el sistema los reconozca como tales (de lo contrario son ignorados). Así para el ejemplo anterior se realizará la búsqueda “hemoglobin AND sequence”. Anotar el número de datos encontrados y repetir la búsqueda utilizando ahora los operadores anteriores ¿Cuántos resultados son obtenidos en cada uno de los casos? ¿Alguna de las búsquedas resultó equivalente a la que se realizó sin empleo de operadores lógicos? ¿Por qué existe una diferencia en el número depublicaciones encontradas con los tres operadores?

3. Para refinar las búsquedas se puede emplear una herramienta del NCBI para construir búsquedas más eficientes. Se puede tener acceso a dicha herramienta mediante el link “Advanced”, localizado debajo de la línea de búsqueda. Construya la sentencia buscando la palabra “hemoglobin” en el titulo y agregar “sequence” en un línea adicional, para buscar también en el título. Observe que puede seleccionar el operador lógico para unir ambos términos. Al llevar a cabo la búsqueda el sistema crea la consulta: “hemoglobin [title] AND sequence [title]”. Compare el resultado obtenido con los resultados de las búsquedas anteriores.

¿Cuál de las búsquedas ha resultado ser la más eficiente hasta el momento?. Explique su respuesta.

4. Modifique la búsqueda anterior, agregando ahora el nombre del autor “Hill RJ” seleccionado el campo “Author” y unir a la sentencia anterior mediante el operador AND. ¿Cuantas publicaciones fueron encontradas?

5. Las consultas se pueden escribir directamente en la línea de búsqueda si se conocen los nombres de los campos deseados en la base de datos. Observe que dichos campos se escriben entre minúsculas inmediatamente después del término buscado. Se pueden utilizar paréntesis para construir sentencias lógicas más complejas. Por ejemplo, realice la búsqueda con la sentencia “(hemoglobin [title] AND sequence [title]) OR (glucose [title] AND oxidase[title])”. Observe los resultados obtenidos y compare con los obtenidos en búsquedas anteriores.

6. También puede observar que PudMed admite otro tipo de opciones de refinamiento de la búsqueda. Dichas herramientas se encuentran actualmente en la parte derecha de la página y se denominan filtros. ¿Qué tipo de refinamiento se puede realizar con dichas herramientas ?

7. Practique con la herramienta ¿puede refinar los resultados para encontrar artículos de una revista particular (p.ej American Journal of Hematology) o publicados en una fecha específica? Cuantos articulos encontro?

II.Consulta de secuencias biológicas mediante ENTREZ.

1. En la página del NCBI se seleccionará ahora de la lista desplegable, la opción “Proteins”. Realizar primero una búsqueda con las consulta “hemoglobin homo sapiens” y observar estos resultados. Posteriormente utilice las opciones de búsqueda avanzada construir la consulta: “(hemoglobin[Title]) AND homo sapiens[Organism]” realice la búsqueda. Compare los resultados de ambas búsquedas ¿cuál de ellas ha resultado ser la más adecuada?

2. Supongamos que está interesado en consultar la secuencia de la cadena alfa de la hemoglobina en el humano. Observe los resultados de la última consulta ¿Cuántos registros para la cadena alfa puede observar? Cuando aparecen varios registros repetidos para una misma molécula nos encontramos ante un caso de "redundancia" de la base de datos. Este fenómeno puede deberse al hecho de que la misma molécula esté depositada en distintas bases de datos (genpept, refseq, uniprot, pdb, etc). Observe que se puede tener acceso a registros de una base de datos determinada aplicado la herramienta de filtros. También con esta herramienta se pueden limitar los resultados para observar solamente los que pertenecen a un organismo particular o bien los que se han liberado en una fecha específica.

3. Ante la redundancia de datos a menudo es conveniente verificar si existen secuencias de referencia. RefSeq es una base de datos del NCBI que organiza las secuencias de referencia en un forma que permite localizar de manera sencilla secuencias asociadas con organismos específicos. A través de RefSeq se pueden localizar moléculas organizadas en torno a genomas ensamblados, de tal forma que a partir de un registro puede tenerse acceso a secuencias de la proteína, mensajeros o secuencia del DNA genómico. Aplique el filtro de RefSeq en la búsqueda para observar solamente los registros pertenecientes a esta base de datos.

¿Cuántas secuencias se reportan en este caso para la hemoglobina?

4. Ingrese ahora a uno de los registros de la hemoglobina alfa que se han encontrado. Observe detalladamente la estructura de este registro e identifique las siguientes secciones: Display,Locus, Accession, Version, Definition, Organism, Comment, Features, Origin. Anote en su reporte sobre el significado de cada una de estas secciones.

Observe que hay un link denominado "GenBank" el cual le informa que está observando el resultado en un formato denominado "GenBank". Este mismo link le permite seleccionar otros tipos de formatos (Summary, FASTA,ASN.1, etc). Observe también que hay un link denominado "Send to" en la parte superior de la página. Este link debe seleccionarse para descargar los registros. En tal caso se selección la opción "file", la cual permite guardar en registros en el formato de texto.

5. Cambie el formato del registro a "FASTA" y observe el cambio en el registro obtenido. Descargar la secuencia obtenida en una carpeta. Al asignar nombres a los archivos no se recomienda modificar las extensiones, ya que a menudo esta extensión la utilizan diversas herramientas informáticas para identificar el formato de los archivos. Nota: Es posible que deba configurar su navegador de Internet para poder elegir la carpeta para almacenar sus archivos.

6. Observe que las claves de acceso o el gi pueden emplearse para buscar un registro específico ¿cuál es la diferencia entre la clave de acceso y el gi que utiliza el NCBI?

7. Realice ahora una búsqueda en la base de datos Nucleotide con la consulta "(hemoglobin[title]AND Homo sapiens [Organism])". Observe que los resultados generados por esta herramienta pueden ser aún más complejos de interpretar que en el caso anterior de para las proteínas.

8. Una base de datos del NCBI muy conveniente para buscar secuencias organizadas en forma biológicamente intituitiva es la base de datos "Gene". Seleccione esta base de datos Gene y realice la búsqueda con la siguiente sentencia: "(((alpha[Protein Name]) AND hemoglobin[Protein Name]) AND Homo sapiens[Organism])" En el listado obtenido ubique los genes correspondientes a las cadenas alfa. ¿Porqué hay dos registros para este gen? Visite el registro de uno de los genes e identifique lo siguiente:
i. El símbolo oficial del gen.

ii. Otras bases de datos donde se le puede localizar.
iii.El cromosoma y la región cromosómica donde se localiza el gen.
iv.Las versiones del ensamblado del genoma humano en las que se basa la secuencia de este gen.
v.El Mapa del gen mostrando la estructura del gen, en la que pueden identificarse (entre otras), exones, intrones, regiones codificantes (CDS), variaciones genéticas (SNPs), datosde ensamblado, datos de expresión (RNAseq).
vi.Exones, intrones, posición del gen en el cromosoma, posición de intrones y exones, localización de regiones codificantes y no codificantes.
vii.Claves de acceso de la proteína, el RNA mensajero, claves de acceso relativas al ensamblado.
viii.Referencias bibliográficas relevantes.
ix.Referencias bibliográficas relacionadas con la función del gen (GeneRIFs: Gene References Into Functions).x.Fenotipos asociados con enfermedades humanas.
xi.Variaciones del gen (ClinVar, dbVar, SNPs).

9. Exploración del genoma humano. Cuando se buscan datos de moléculas provenientes de un genoma ya secuenciado se pueden utilizar los navegadores de genomas (Genome browsers) para explorar detalladamente la ubicación de los genes. En la sección de genomas del NCBI (http://www.ncbi.nlm.nih.gov/Genome/) se muestra un resumen de los proyectos de secuenciación los organismos disponibles en la actualidad. En “Genome resources” se muestra una clasificación de los proyectos en función del tipo de organismo. Tambien se muestran algunos enlaces para recursos disponibles del genoma , búsquedas con Blast, mapas genómicos y páginas principales del proyecto de secuenciación para algunos organismos importantes. Buscar “Human” y entrar a la opción del mapa genómico de este organismo. Se muestra una página en la que se representan los cromosomas del genoma humano y una sección para búsquedas. Activar la sección de “Advanced search” y en la página de búsqueda introducir la sentencia de búsqueda “HBB OR HBA2”. En la sección “Type of mapped object” desmarcar todas las opciones excepto “Gene” y en “Assembly” seleccionar “Reference”. Presionar el botón “Find” y observar la estructura de los resultados. Observar las marcas rojas en el mapa que muestran la ubicación de los elementos encontrados en la búsqueda y compárelas con los datos de la tabla de resultados. Presionar en el elemento de mapa “HBB” y observar los resultados. Se muestra un acercamiento a la región cromosómica analizada, destacando la zona donde se encuentra el elemento buscado. La página tiene opciones para realizar acercamientos a la región de interés (Zoom). Presionando sobre el símbolo oficial del elemento buscado (HBB) la página nos dirige a la sección “Gene” de este gen que se describió en el punto anterior. ¿Cuáles son los elementos cromosómicos anotados más cercanos al gen HBB? Repita este análisis para el gen de la alfa globina.

EJERCICIOS DE EVALUACION.
Elaborar un reporte de texto con la siguiente información:
E1. Las secuencias de proteínas en formato FASTA de las cadenas alfa y beta de la hemoglobina humana descargadas en esta práctica y las secuencias de nucleótidos de los mRNA completas que las codifican también en formato FASTA.

E2. Elaborar en cuadro en el cual se resuma la siguiente información para las proteínas estudidadas: Clave de acceso del cromosoma, la proteína y el mRNA; número de intrones y exones, longitud del gen, longitud de la proteína, longitud del mRNA, posiciones en el cromosoma de cada gen, posiciones de los intrones y posiciones de las regiones UTR.

E3. En un cuadro enumerar las diferencias existentes entre la anotación de los genes de la hemoglobina alfa y beta en el NCBI y en Ensembl.

E4. Utilizando la sección de genomas del GenBank realizar la búsqueda de genes para 16S rRNA´s en el genoma de Escherichia coli K-12.

PREGUNTAS (CONTESTAR BREVEMENTE):
1. ¿Cuántos genomas eucarióticos y procarióticos se han secuenciado de manera completa en la actualidad?
2. Explique el concepto de redundacia de la bases de datos y explique en que consiste la base de datos RefSeq del NCBI.
3. Investigue en que consiste la base de datos Uniprot.
4. ¿Cuál es la diferencia entre la clave de acceso (accession) y el GenInfo Identifier (gi) que aparece en los registros de secuencias del NCBI? ¿Cuál es la utilidad de estas claves?

Referencias bibliográficas.
1. Fernández-Suárez XM, Rigden DJ, Galperin MY. (2014): The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res. 42(Database issue):D1-6.
2. niProt Consortium1. (2007): The Universal Protein Resource (UniProt). Nucleic Acids Res. 35 (Database issue):D193-7.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2007): GenBank. Nucleic Acids Res. 35(Database issue):D21-5.
4. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R. (2007): EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res. 35(Database issue):D16-20.
5. Sugawara H, Abe T, Gojobori T, Tateno Y. (2007): DDBJ working on evaluation and classification of bacterial genes in INSDC. Nucleic Acids Res. 35(Database issue):D13-5.
6. Berman H, Henrick K, Nakamura H, Markley JL. (2007): The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35(Database issue):D301-3.
7. NCBI staff (2002-2005): The NCBI Handbook, McEntyre, J.; Ostell, J., editors Bethesda (MD): National Library of Medicine (US), NCBI; http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.TOC&depth=2
8. Claverie JM, Notredame C. (2003): Bioinformatics for dummies. Wiley, New York,USA, pp: 73-173.
9. Gibas C, Jambek P. (1999): Developing bioinformatics computer skills. O´Reilly, USA, pp: 131-156.