PRACTICA II: 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.
- Buscar y descargar secuencias biológicas en otras bases de datos primarias.
- Conocer como obtener estructuras de proteínas almacenadas en la base de datos PDB.

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 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 estas en general 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 denamonidas 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 2012 contiene una lista de 1380 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/Welcome-e.html (DNA Databank of Japan- DDBJ)
http://www-nbrf.georgetown.edu/pirwww/ (Protein Information Resource- PIR)
http://us.expasy.org/sprot/ (Swiss-Prot)
http://www.rcsb.org/pdb/home/home.do (Protein Data Bank- PDB)

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 barra en color azul obscuro 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 presionando el botón "go" mostrado en la página principal junto a la línea para búsquedas). Presionando PubMed en la barra o en la lista detallada 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.

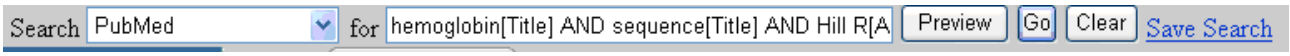
1. En la línea para búsquedas teclear la frase "**hemoglobin sequence**" y presionar el botón "**Go**". Observar el número de referencias obtenidas y la forma como son clasificadas por el sistema. Examine algunos de los títulos de los trabajos encontrados ¿Considera apropiado el número de referencias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

2. Realizar la búsqueda anterior pero utilizando los 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 de publicaciones encontradas con los tres operadores?

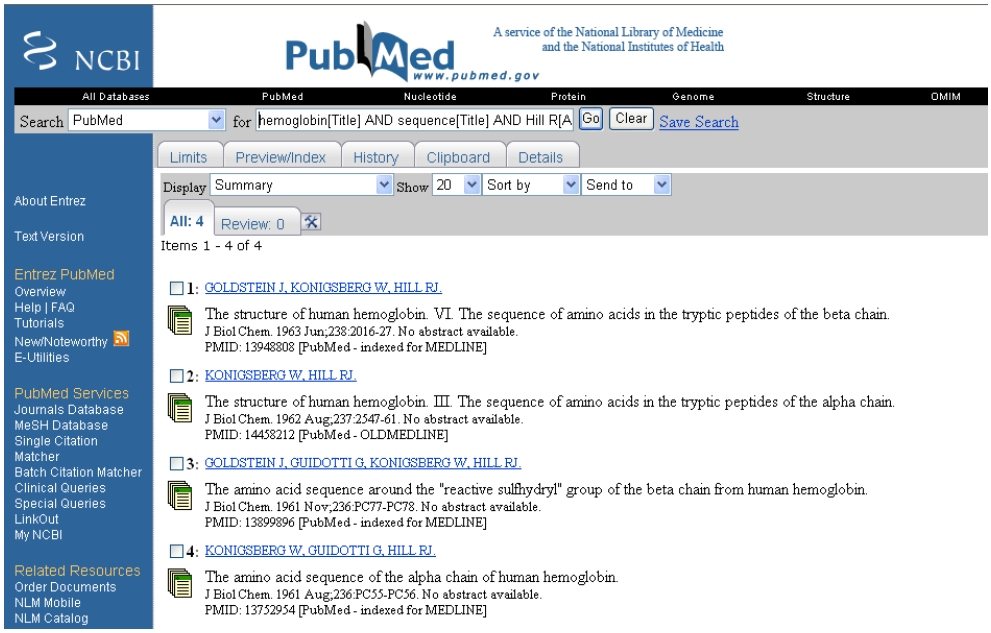
3. Realizar la búsqueda anterior pero ahora escribiendo la sentencia "**hemoglobin [title] AND sequence [title]**" y comparar el resultado obtenido con los resultados de las búsquedas

anteriores. Observe los títulos de las referencias encontradas ¿Qué diferencia principal tienen los resultados obtenidos en esta búsqueda con respecto a los anteriores?

4. Los nombres entre corchetes corresponden a los campos de datos que contienen los registros en las bases de datos y dependiendo de cual es consultada se dispone de diferentes campos. Los campos permiten restringir las búsquedas con mayor eficiencia para disminuir el número de resultados indeseables. Adicionalmente el NCBI dispone de una herramienta que permite seleccionar fácilmente cada uno de los términos deseados para elaborar la consulta (query). Para emplearla dar un clic en la sección "Preview/index", después en la parte denominada "*Add Term(s) to Query or View Index*" escribir en la línea de texto el término "hemoglobin" y seleccionar el campo "Title" desplegando el menú existente. Después presionar el botón "Preview", para observar el resultado de la búsqueda. Escribir ahora "sequence" y seleccionar el campo "Title", presionar entonces el botón del operador "AND". Ahora presione "Preview" y observe el resumen del resultado parcial. Nótese que en la línea de búsqueda aparece la construcción actual. Presionando "Go" se obtienen los resultados de la misma. Agregar ahora el nombre del autor "Hill R" seleccionado el campo "Author" y unir a la sentencia anterior mediante el operador AND. Presionar "Preview" ¿Cuantas publicaciones fueron encontradas?



Para revisar las publicaciones obtenidas bastara hacer clic en el número de resultados obtenidos "Result", actualmente 4, el sistema nos mostrará un listado con las cuatro publicaciones encontradas con nuestra búsqueda. Observe que estas publicaciones tienen un icono de archivo con una franja en color verde, esto significa que el artículo completo es gratuito. Al hacer clic en el icono se abrirá otra ventana donde se mostrará el resumen del artículo y un link para la descarga del mismo. Nota: Algunas referencias vienen marcadas con iconos de distintos colores. Estos códigos indican que el artículo completo es libre pero que se puede descargar del sitio de la revista (icono anaranjado), de la base de datos del NCBI llamada PMC (icono verde) o de ambos (icono verde y anaranjado).



5. Se pueden utilizar paréntesis para construir sentencias lógicas más complejas. 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. PudMed admite otro tipo de opciones de refinamiento de la búsqueda. Para ilustrar su funcionamiento escribir la sentencia de búsqueda "**hemoglobin [title] AND sequence [title]**" y después dar un clic en la opción "LIMITS". Se muestra un cuadro con diversas opciones de configuración. Buscar la sección "Dates" y en "Published in the Last:" seleccionar de la lista de opciones "Specify data range "YYYY/MM/DD)" Se activa una sección en la que se solicita un intervalo de fechas, escribir 2000 a 2002 en los cuadros correspondientes y presionar el botón "Go". Observe las fechas de los resultados obtenidos. ¿Qué otras opciones en "LIMITS" están disponibles para el refinamiento de resultados?

II. Consulta de secuencias biológicas mediante ENTREZ y RefSeq.

En ENTREZ se tienen distintas bases de datos mediante las cuales se puede tener acceso a secuencias biológicas. Tradicionalmente, las base de datos "protein" and "nucleic" se han empleado para buscar secuencias de aminoácidos o nucleótidos respectivamente. No obstante, ante la enorme cantidad de datos que se han acumulado recientemente y la redundancia de muchos de ellos, se ha puesto especial interés, en organizar dicha información en torno a genomas modelo o de referencia. La base de datos de secuencias de Referencia del NCBI es una colección de registros de secuencias genómicas, transcritos y proteínas. Estos registros se seleccionaron y curaron a partir de archivos de secuencias públicas, para disminuir su redundancia. La secuencias se organizan en torno a organismos modelo, procurando tener un único registro la secuencia del genoma, genes (y sus isoformas) y sus productos proteicos. Las formas más apropiadas para buscar información en esta base de datos es mediante el acceso a los mapas genómicos de los organismos modelo, o bien a la base de datos "Gene" del NCBI.

1. Para este ejercicio en ENTREZ puede seleccionar la base de datos "Gene" y escriba la siguiente instrucción para realizar una consulta:

   hemoglobin [title] **AND** homo sapiens [organism]

   Observe la lista de registros que se obtienen con esta consulta. Identifique los registros que corresponden a los genes de la alfa y beta globinas. ¿Cuántos registros identifica para estos genes? ¿en que cromosoma y región se localizan estos genes? ¿cuáles son los símbolos oficiales que se emplean para representarlos? Identifique otros genes correspondientes a otras hemoglobinas (hemoglobinas fetales) ¿en que cromosomas y regiones se localizan? Identifique finalmente pseudogenes de las alfa y beta globinas y en donde se localizan.

2. Ingrese al registro de la beta globina humana (HBB). Analice cuidadosamente la información en este registro y trate de identificar la siguiente información: a) Cromosoma y región cromosómica en la que se localiza el gen b) Intervalo de nucleótidos con respecto al ensamblado primario del genoma humano (GRCh37.p9) c) Representación gráfica del gen d) Claves de acceso del genoma (cromosoma), RNA mensajero y Proteína e) ¿existen isoformas de este gen? f) secuencias de nucleótidos y de aminoácidos del gen, RNA mensajero y proteína en los formatos GenBank y FASTA, g) Posiciones en el genoma y RNA mensajero de intrones, exones, regiones codificantes (CDS), regiones no codificantes (UTRs, Untranslated Regions).

3. Al ingresar al registro de una secuencia, por ejemplo, el de la proteína alfa globina, encontrará un registro con una amplia variedad de información. Dicha información, constituye la anotación del registro, que puede dividirse en a) aquella empleada para clasificar e identificar un registro

específico en la base de datos b) la anotación biológica de la secuencia. Los registros incluyen claves de acceso, número de versión, identificadores de GenBank o GenPept, clasificación taxonómico y referencias bibliográficas, para lo referente a información para clasificación. Dentro de la información biológica pueden encontrarse datos respecto a la función, anotación sobre funciones específicas de regiones de la secuencia y desde luego, la secuencia de la molécula. Además observe que hay distintos formatos para visualizar los registros. Los más comunes son el formato GenBank/GenPept (detallado) y el formato FASTA (reducido). Adicionalmente la página incluye herramientas para descargar secuencias y modificar la presentación de la secuencia, ajustando sus límites, intercambiar entre cadena directa y complementaria, etc. Finalmente, se tienen links hacia otras herramientas con las cuales pueden realizarse análisis específicos con la secuencia, tal como BLAST.

4. 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/Genomes/) 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. En "Organism-specific" se muestran algunos enlaces para recursos disponibles del genoma (G), búsquedas con Blast (B), mapas genómicos (M) y páginas principales del proyecto de secuenciación (P) 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.

III. Consulta de secuencias en otras bases de datos biológicas mediante SRS.

La herramienta Sequence Retreival System (SRS) es un sistema de búsqueda similar a Entrez que utilizan comúnmente otras bases de datos tales como EBI-EMBL, SWISS-PROT y la DDBJ aunque la interfaz de acceso muestra algunas diferencias entre estas bases de datos. A manera de ejemplo localizaremos las secuencias de aminoácidos previamente estudiadas de la alfa y beta globinas humanas en la base de datos del EMBL.

1. Se puede tener acceso al sistema SRS desde diversas partes del sitio del EBI-EMBL pero la siguiente dirección nos remite directamente a este sistema (http://srs.ebi.ac.uk/). En la página mostrada seleccionar "*Protein*" en la sección "*Find*" y después activar la carpeta en la parte superior denominada "*Library page*". En esta página se deben seleccionar las bases de datos que se desea consultar. Buscar la sección "*Uniprot Universal Protein Resource*" y marcar la opción "*UniProtKB/Swiss-Prot*" que corresponde a la base de datos depurada del SWISS-PROT.

2. Activar la carpeta "*Query Form*" en la parte superior de la página con lo cual se muestra un cuadro para la configuración de las búsquedas (En esta página está disponible tambíen una opción para búsquedas avanzadas "*Extended Query*"). En la sección "*Fields you can search*" escribir "*beta globin*" y seleccionar de la lista desplegable "*Reference: Title*". En la segunda línea escribir "*Homo sapiens* " y seleccionar de la lista "*Organism name*". Verificar en "*Search options*" el operador empleado para combinar los criterios, el cual debe ser "& (AND)" y presionar el botón "*Search*".

3. En el cuadro de resultados buscar el registro HBB_HUMAN y entrar en él. Observe el registro extensamente detallado para esta proteína. En la parte superior del registro existen enlaces que le permiten dirigirse a alguna sección determinada del registro tal como la secuencia. Del lado izquierdo hay un cuadro de opciones ("*Entry options*") el cual tiene un botón "*Save*". Presionar este botón y activar las opciones necesarias para grabar el registro como archivo de texto en formato FASTA (FastaSeqs). Repetir la operación para buscar la secuencia de la proteína de la alfa globina humana.

IV. Descarga de estructuras.

La base de datos del Protein Data Bank almacena datos de coordenadas tridimensionales para estructuras de proteínas determinadas por cristalografía de difracción de rayos X o por resonancia magnética nuclear. Se puede tener acceso a la misma presionando el link http://www.pdb.org/pdb/home/home.do.

1. Al ingresar a esta página, en la parte superior hay una línea de búsqueda en la cual se puede escribir él número de acceso de la estructura (PDB ID) o el nombre de la proteína. También hay un enlace para realizar búsquedas avanzadas.

2. Para este ejemplo escribir la clave de acceso 1DXT y presionar el botón "Site search". Se muestra un cuadro de resultados relacionado con esta estructura. En la parte superior junto al ID de la estructura hay un icono que permite descargar la estructura.

3. Almacenar la estructura en la carpeta "c:\\Bioinfo\estruct" con el nombre 1DXT.pdb y posteriormente edítela con un editor de texto. Identificar detalles importantes en el registro tales como el título de la estructura, la resolución, la secuencia de aminoácidos y las coordenadas de cada uno de los átomos. Para visualizar este tipo de archivos se requiere un programa visualizador de estructuras como se estudiará en la siguiente práctica.

GUÍA PARA EL REPORTE DE LA PRÁCTICA.

1. 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.

2. Investigar el polimorfismo de nucleótido sencillo (SNP, Single Nucleotide Polymorphism) en el gen de la hemoglobina beta, asociado al anemia de células falciformes. Localice el registro de NCBI de este polimorfismo, identifique la variación, la forma de reportarla e investigue si puede mostrar su ubicación empleando las herramientas de navegación gráfica del NCBI.

3. Utilizando la base de datos del NCBI realizar la búsqueda de genes para 16S rRNA´s en el

genoma de *Escherichia coli* K-12. ¿De que tamaño es el genoma de esta bacteria? ¿Cuántos genes para 16S rRNA se localizan? ¿Qué diferencia existe en la organización de registros en el NCBI para organismos procarióticos y eucarióticos?

4. Realizar la búsqueda de la beta globina humana en las bases de datos SWISS-PROT y DDBJ.

5. La base de datos PIR tiene un sistema de búsqueda diferente a Entrez y SRS, localice la secuencia de la beta globina humana en esta base de datos y reporte algunas de las claves de acceso encontradas para esta proteína.

PREGUNTAS EXTRA (CONTESTAR BREVEMENTE):

1. ¿Cuántos genomas eucarióticos y procarióticos se han secuenciado de manera completa en la actualidad? Emplee la información del NCBI para responderlo.

2. Explique el concepto de redundancia 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?

5. Una base de datos muy empleada en la actualidad es ENSEMBL, mantenida por el EMBL. Investigue en qué consiste y cuales son las diferencias más importantes entre esta base de datos y la RefSeq del NCBI.

Referencias bibliográficas.

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2007): *GenBank*. Nucleic Acids Res. 35(Database issue):D21-5.
2. 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.
3. Claverie JM, Notredame C. (2003): *Bioinformatics for dummies*. Wiley, New York,USA, pp: 73-173.
4. Galperin MY. (2007): *The Molecular Biology Database Collection: 2007 update*. Nucleic Acids Res. 35(Database issue):D3-4.
5. Gibas C, Jambek P. (1999): *Developing bioinformatics computer skills*. O´Reilly, USA, pp: 131-156.
6. 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.
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. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. (Database issue), pp: D130-D150.
9. 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),pp:D13-D15.
10. UniProt Consortium1. (2007): *The Universal Protein Resource (UniProt)*. Nucleic Acids Res. 35 (Database issue)pp:D193-D197.