Taller01: Bases de Datos Biológicas: Interacción con algunas BDs del NCBI

6 de agosto de 2018

1. Objetivo

Las siguientes prácticas tienen por objeto aprender a manejar la información contenida en el NCBI de una forma simple.

2. Introducción

A lo largo de los últimos 15 o 20 años, se ha ido acumulando una gran cantidad de información de naturaleza molecular (secuencias de genes , genomas, proteínas, etc.) , procedente de los distintos proyectos genoma de diferentes especies (Homo sapiens, Pantroglodytes, Gallus gallus, Drosophila melanogaster, Takifugu rubripes, Caenorhabditis elegans, entre otros.).

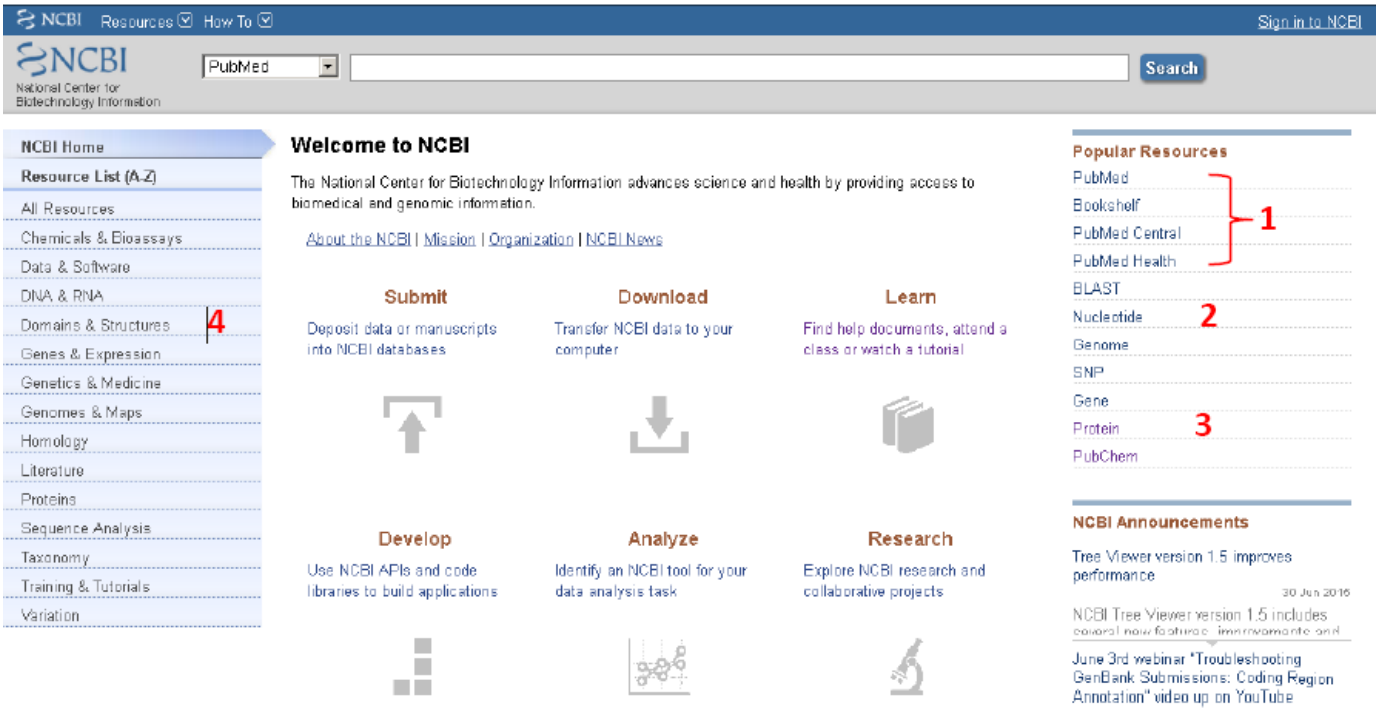
Toda esta información se ha ido depositando en grandes "almacenes" de información de secuencias, organizadas en bases de datos, con la intención de que científicos y público en general, pudiera acceder a ella a través de internet. Como complemento a esa información de tipo molecular, estos "almacenes" han incorporado toda una colección de publicaciones y textos científicos de tipo biomédico. En este sentido, el que un biólogo sepa cómo acceder y explotar esta información de un modo eficiente, resulta hoy en día algo absolutamente imprescindible y necesario.

De todos estos almacenes de información de secuencias, el correspondiente al "National Center for Biotechnology Information (NCBI)" puede considerarse como el de referencia en lo que a obtención de secuencias moleculares y publicaciones biomédicas se refiere.

3. Guía

3.1. Ingreso al sitio del NCBI

Ingrese al sitio del NCBI, la URL (Uniform Resource Locator) del NCBI es `http://www.ncbi.nlm.nih.gov`, y su página inicial es a día de hoy la siguiente:



En la gráfica hemos indicado los enlaces que nos llevan a los contenidos de información relativos a:

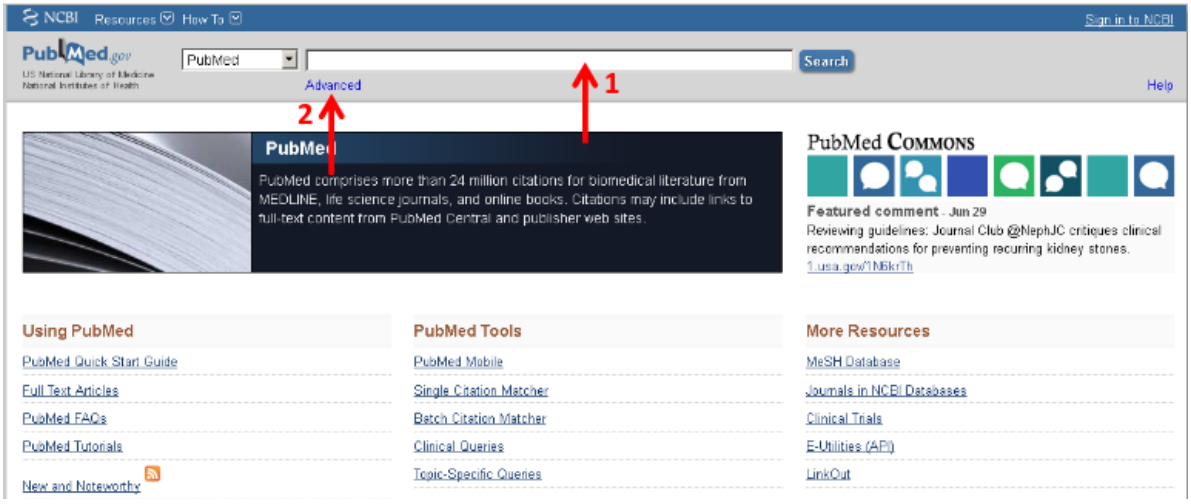
- Publicaciones de índole biomédica (1),
- De secuencias de nucleótidos (2)
- De secuencias de proteínas (3), y
- De la estructura tridimensional de moléculas (4) .

3.2. Enlaces a publicaciones relacionadas con literatura biomédica

- **Pubmed** : PubMed comprende más de 24 millones de citas de la literatura biomédica , revistas de ciencias biológicas, y los libros en línea. Las citas pueden incluir vínculos al texto completo de artículos de PubMed Central (ver más abajo) y sitios web de editoriales, o solamente al resumen de dichos artículos.
- **Bookshelf** : Proporciona acceso gratuito a textos en línea y documentos en ciencias de la vida y de la salud .
- **PubMed Central** : Es un archivo de revistas de carácter biológico y biomédico, de libre acceso, y depositado en la Biblioteca Nacional de Medicina, de los Institutos Nacionales de Salud (NIH/NLM).
- **PubMed Health** : Proporciona información a médicos y público en general sobre la prevención y tratamiento de enfermedades y afecciones.

Guía:

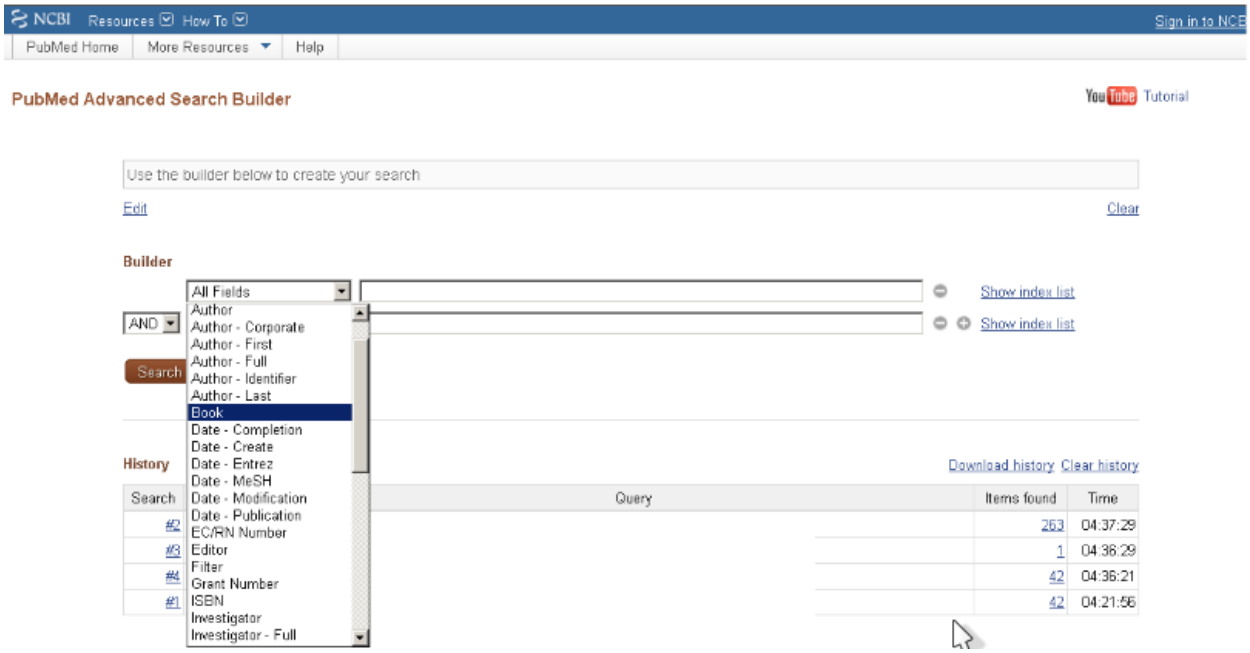
1. Vamos a buscar referencias biomédicas a través de PubMed sobre la organización del promotor de eucariotas. El punto de partida de la búsqueda puede realizarse desde distintos sitios, pero para sistematizar este procedimiento, vamos a realizar la búsqueda desde la página inicial de PubMed. Para ello seleccione (click) en el enlace PubMed que vemos en la figura de más arriba, situado en la columna encabezada por “Popular Resources”, lo que nos lleva a la siguiente página:



2. En la ventana de búsqueda (señalada con una flecha 1) podemos incluir los términos de búsqueda (generalmente, en inglés): "eukaryotic promoter organization", lo que nos da una relación de cerca de más de 300 artículos en los que aparecen cualquiera de los términos anteriores, que posteriormente podemos reordenar de acuerdo a distintos criterios: relevancia, tipo de artículo (revisiones, descripciones completas de un paciente o enfermedad - “case report” - , carta, noticia, etc.) , periodo de publicación en años, etc.

3. La búsqueda le presenta dos tipos de organización de los resultados: por relevancia (*best match*) o por tiempo de publicación *(Most recent),* siendo por defecto (por ahora) la organización por tiempo de publicación. Seleccione de cada uno de estos tipos un artículo (cualquiera) y en sus palabras describa de que puede tratar el artículo o a que se refiere la investigación descrita en el mismo.

4. Alternativamente, podemos realizar una búsqueda avanzada de artículos (señalada con la flecha 2), en la que podemos incluir términos específicos para campos concretos de la base de datos de PubMed para hacer la búsqueda más específica y precisa. Algunos de estos términos son: autor, fecha de publicación, idioma de la publicación, revista, entre otros. Con estos criterios realizar una búsqueda de publicaciones entre loa años 2000 y 2005 y seleccione un artículo y describa de que posiblemente trata la investigación.



La búsqueda de información en las restantes bases de datos PubMed Central, Bookshelf o PubMed Health, es similar a lo mostrado anteriormente.

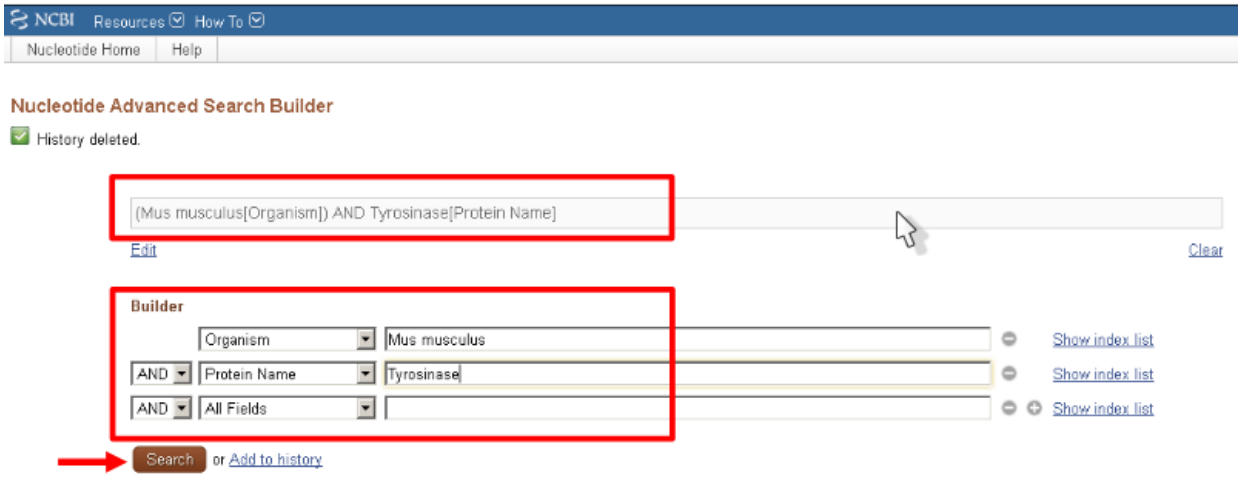
3.3. Búsqueda y obtención de secuencias de nucleótidos

El procedimiento es muy similar al indicado para buscar información en PubMed, sólo que ahora trabajaremos en una base de datos del NCBI diferente; en este caso será la base datos de "Nucleotide".

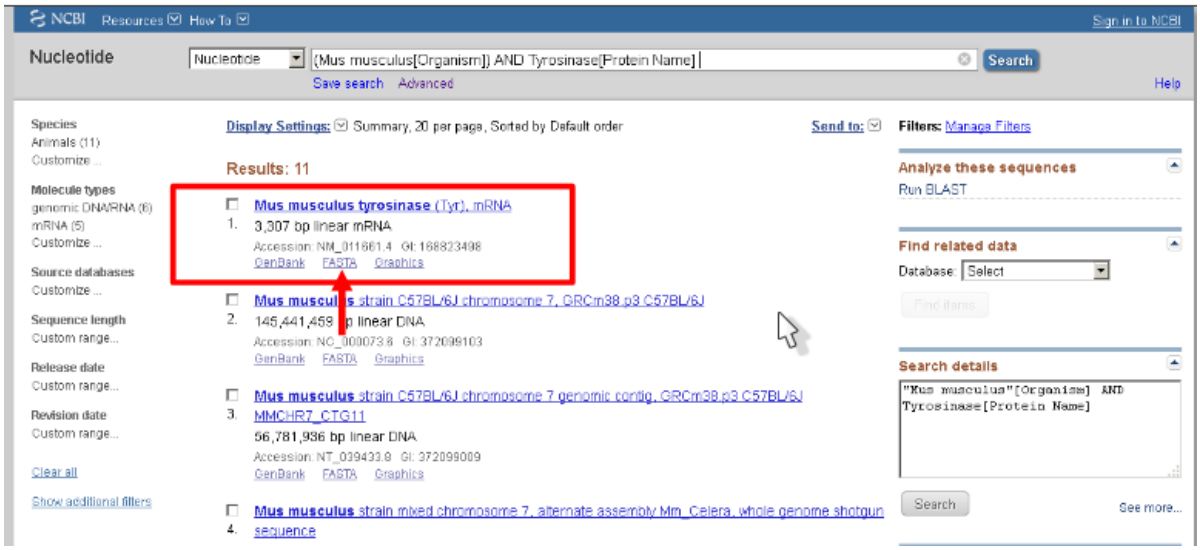
1. En la página principal de NCBI seleccionamos el enlace correspondiente a Nucleotide ("Popular resources", columna de la derecha), y entramos en la página inicial de NUCLEOTIDE.

   Al igual que veíamos en PubMed, podemos introducir los términos de búsqueda, bien la ventanita (flecha 1) o bien a través del procedimiento de búsqueda avanzada (flecha 2). Esto último es generalmente preferible, puesto que podemos afinar mucho más nuestra búsqueda.

2. Aquí vamos a buscar la secuencia del mensajero del gen (mRNA) de la Tirosinasa en el ratón (mutaciones en el gen de la tirosinasa, producen albinismo) usando el procedimiento de búsqueda avanzada. Introduciremos sucesivamente los términos *Mus musculus* y *tyrosinase* en los campos de "*organism*" y "*protein name*"



La respuesta tendría el siguiente aspecto:



3. De este resultado obtenga la secuencia de nucleótidos en formato FASTA y copie en el informe los primeros 700 nucleótidos, junto con el encabezado, todo en formato FASTA.

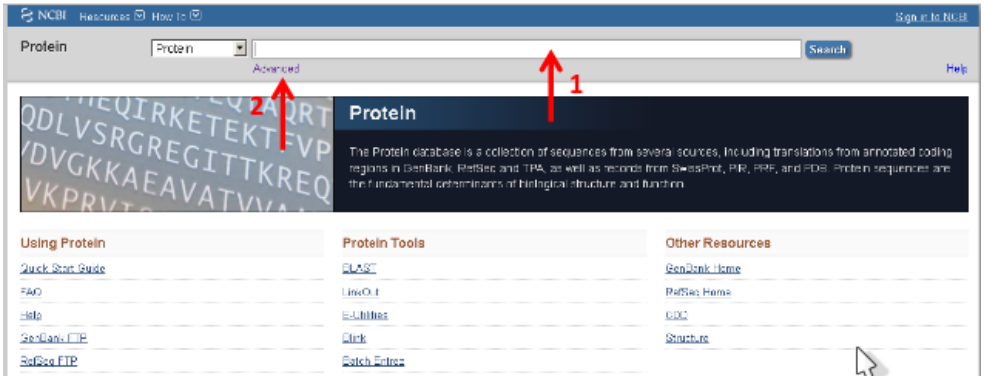
   Una secuencia en formato FASTA, bien de nucleótidos o de aminoácidos, tiene una sintaxis caracterizada por una primera línea que obligatoriamente empieza por el símbolo "mayor que" (>) seguido por una identificación de la secuencia en cuestión; esta línea es meramente informativa. A partir de la segunda línea y

siguientes aparece la secuencia de la molécula propiamente dicha. Por ejemplo, la secuencia de nucleótidos ATTGCCGTTATGCAATTGAT en formato FASTA aparecería como sigue:

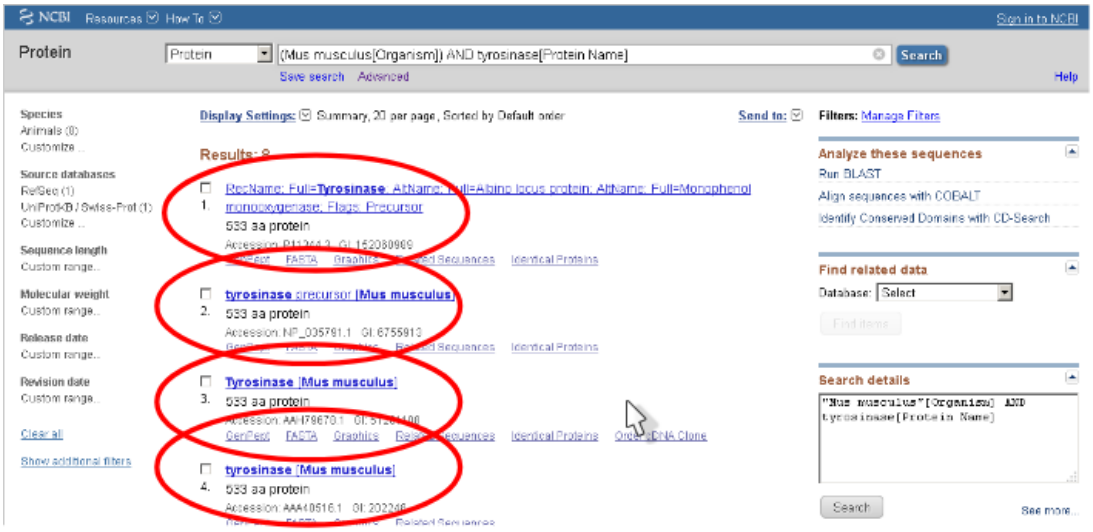
```
>Ejemplo de secuencia en FASTA
ATTGCCGTTATGCAATTGAT
```

3.4. búsqueda y obtención de secuencias de proteínas

El procedimiento de búsqueda es totalmente equiparable al de las búsquedas de secuencias de nucleótidos, sólo que la base de datos del NCBI sobre la que se ha de trabajar es la de “Protein”. Podemos acceder a ella desde la página principal de NCBI; seleccionando el enlace correspondiente a Protein (“Popular resources”, columna de la derecha), y entramos en la página inicial de PROTEIN.



1. Realizar la búsqueda de la secuencia proteica de la tirosinasa (tyrosynase) del ratón (Mus musculus) a través del procedimiento de búsqueda avanzada:



2. Seleccione cualquiera de las primeras entradas y obtener la secuencia de la proteína buscada, copie las 10 primeras líneas de la secuencia en el informe en formato FASTA.

tyrosinase precursor [Mus musculus]

NCBI Reference Sequence: NP_035791.1

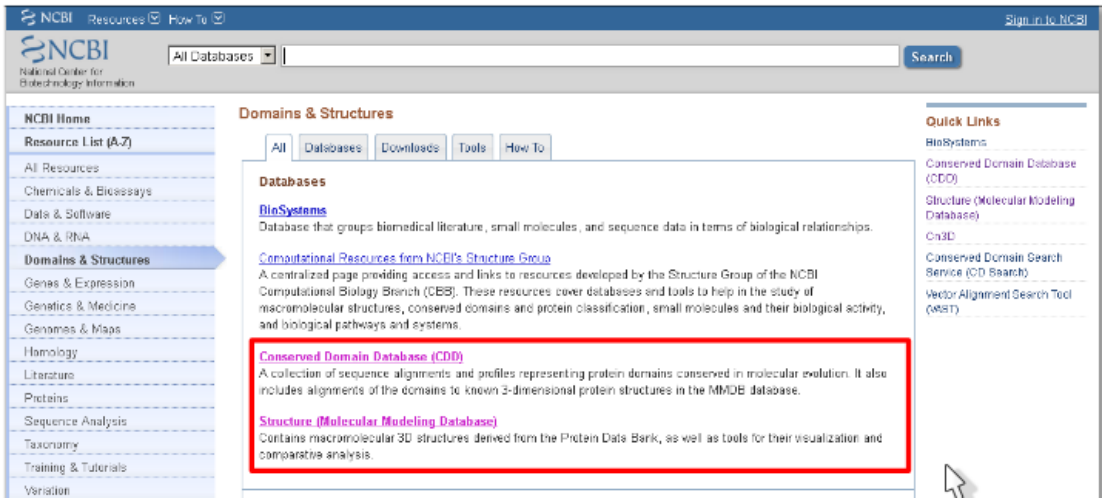
GenPept Identical Proteins Graphics

```
>gi|6755913|ref|NP_035791.1| tyrosinase precursor [Mus musculus]
MFLAVLYCLLWSFQISDGHFPRACASSKNLLAKECCPPWMGDGSPCGQLSGRGSCQDILLSSAPSGPQFP
FKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKRVLIRRNIFDLSVSEKNKFFSYLTL
AKHTISSVYVIPTGTYGQMNNGSTPMFNDINIYDLFVWMHYYVSRDTLLGGSEIWRDIDFAHEAPGFLPW
HRLFLLLWEQEIRELTGDENFTVPYWDWRDAENCDICTDEYLGGRHPENPNLLSPASFFSSWQIICSRSE
EYNSHQVLCDGTPEGPLLRNPGNHDKAKTPRLPSSADVEFCLSLTQYESGSMDRTANFSFRNTLEGFASP
LTGIADPSQSSMHNALHIFMNGTMSQVQGSANDPIFLLHHAFVDSIFEQWLRRHRPLLEVYPEANAPIGH
NRDSYMVPFIPLYRNGDFFITSKDLGYDYSYLQESDPGFYRNYIEPYLEQASRIWPWLLGAALVGAVIAA
ALSGLSSRLCLQKKKKKKQPQEERQPLLMDKDDYHSLLYQSHL
```

3.5. Búsqueda y obtención de estructuras tridimensionales

El punto partida para obtener la estructura tridimensional de macromoléculas es el enlace “Domains & Structures” situado la página principal del NCBI, en la columna de la izquierda.

1. Seleccione el enlace anterior para llegar a la página que nos permite acceder a las bases de datos de estructuras moleculares tridimensionales.



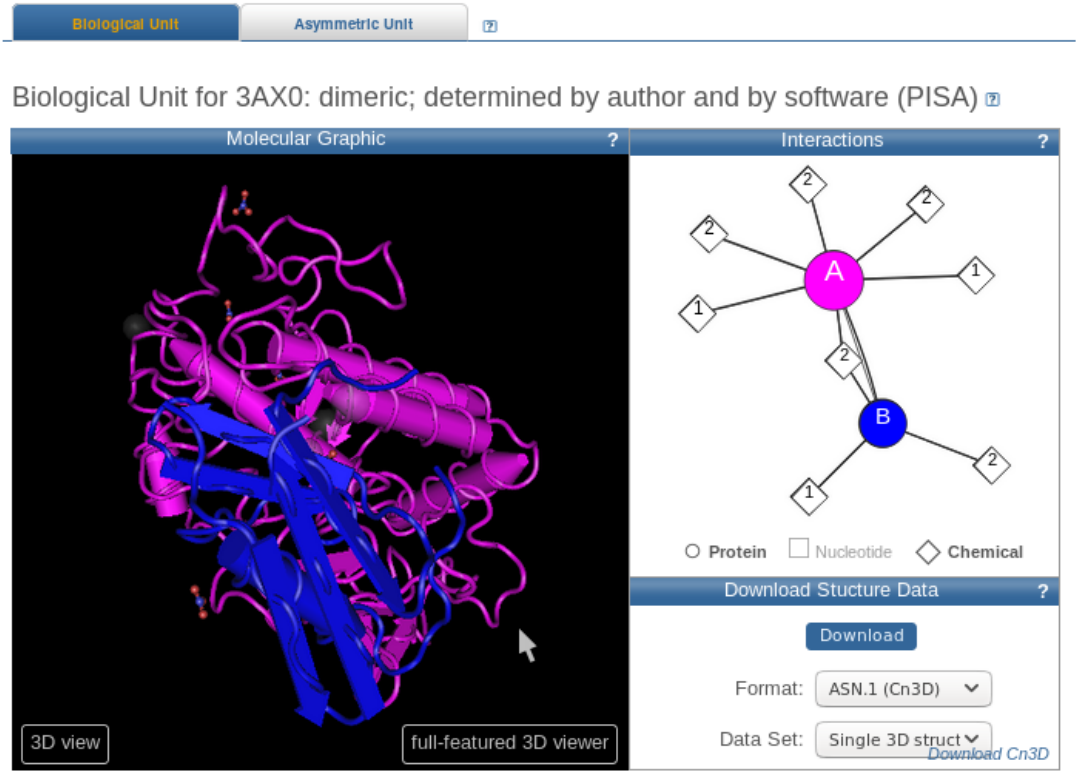
Estas dos bases de datos que vemos recuadradas en la figura, se refieren a la colección de estructuras 3D de una serie de dominios de proteínas conservados a lo largo de la evolución (CDD), y a la colección de estructuras 3D de macromoléculas.

Para buscar información en ellas se operaría exactamente igual que en el caso de PubMed, Nucleotide, y Protein. Por ello, no vamos a hacer ninguna indicación especial en ese sentido.

2. Busqué las estructura 3D de la proteína tirosinasa (*tyrosinase)* en la BD de estructuras *(Structure (Molecular Modeling Database)*

   No obstante, para poder visualizar estas estructuras en modo 3D, se necesitan programas específicos. NCBI utiliza el visualizador Cn3D como estándar, que funciona solamente en Windows, pero existen otras herramientas como ***rasmol*** que funcionan en Linux, Mac y Windows. Sin embargo, el sitio mismo del NCBI ofrece un visualizador de estructuras de proteínas integrado el cual puede seleccionar para visualizar la estructura y sus dominos, entre otros elementos.

3. Seleccione este visualizador y obtenga tanto la imagen espectirosinasaífica de uno de estos dominios junto con su secuencia de aminoácido correspondiente.



4. Descargue el archivo de la estructura 3D anterior (archivo .pdb) y copie las posiciones de los primeros 10 átomos de la proteína.

3.6. Comparación de los diferentes tipos de archivos de las BDs.

Con todo lo anterior, descargue tres tipos de archivo de la proteína tirosinasa (regrese a las búsquedas anteriores y descargue el archivo en formato FASTA, GenBanck, y PDB. Escriba en el informe un pedazo representativo de cada archivo y compare su estructura.

4. INFORME

Realice un informe sobre los puntos que están en letras rojas. No coloque pantallazos, sino el texto o la figura original.

5. Referencias

- Este taller es una versión actualizada y modificada del descrito en: `https://www.uam.es/personal_pdi/ciencias/gpepe/g-molecular/Practicas/GM_guion_practica_bioinfo.pdf`