Fuentes de datos en Bioinformática

Bases de datos biológicas

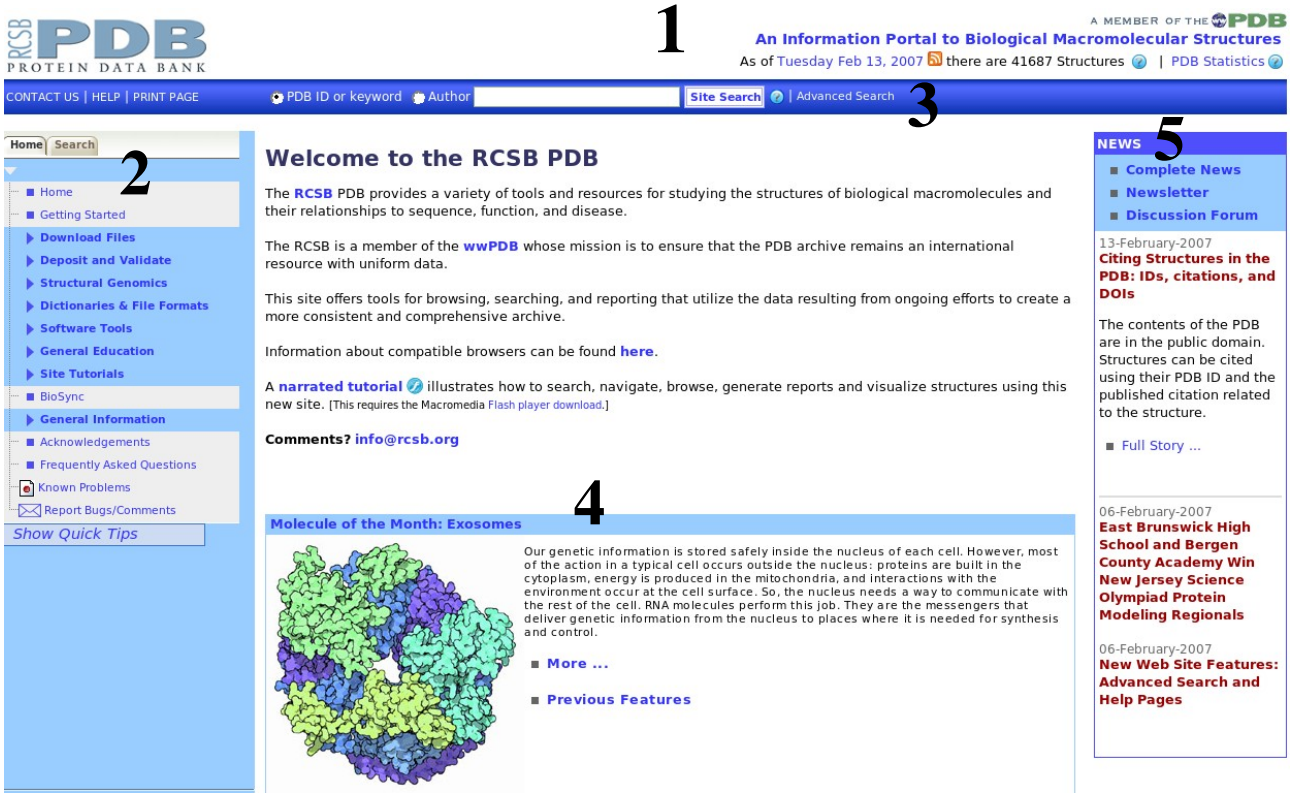
Introducción

Con el desarrollo de esta guía podrá familiarizarse con el uso e interpretación de dos de las bases de datos más importantes en bioinformática: PDB y Uniprot. En sesiones posteriores hablaremos detalladamente de otras bases de datos de gran interés, tales como EMBL y Genbak.

PDB

Visite la siguiente dirección: http://www.rcsb.org/

En este momento se debe encontrar en el sitio web del Protein Data Bank (PDB):



Tal vez encuentre algunas diferencias , de acuerdo a la fecha en la que esté accesando este sitio. Sin embargo siempre encontrará que el sitio web de PDB tiene las misma arquitectura:

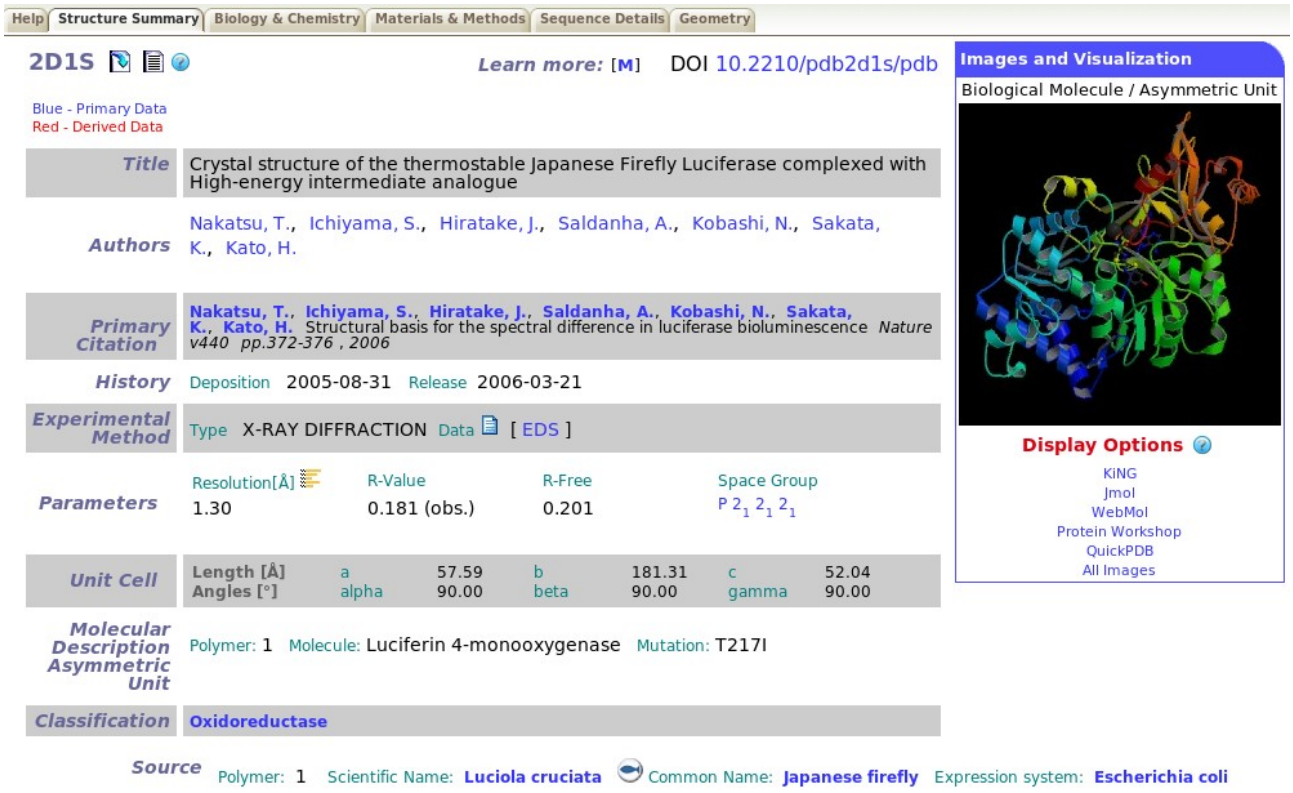
1. Información general de la base de datos, tales como número de registros y estadísticas.
2. Menú de navegación: donde se encuentran las diferentes secciones del sitio. Puede ser de especial ayuda para esta primera aproximación al PDB que consulte el enlace "**gettin Started**".
3. Barra de consulta: por medio de esta barra es posible realizar búsquedas (básicas o avanzadas) de nuestra molécula de interés en PDB, ya sea por ID, palabra clave o autor.

4. Molécula del mes: mensualmente PDB selecciona una molécula, para la cuál provee información estructural y funcional muy completa.
5. Barra lateral de noticias, con información concerniente a PDB.

¿Con cuantos registros cuenta actualmente el PDB?
¿Cual es el total de estructuras obtenidas por cristalografía de rayos X y microscopía electrónica en el PDB hasta este año?

Ingrese el siguiente identificador en la casilla de búsqueda: **2d1s**
Presione el botón "**site search**"

Después de unos segundo se encontrarácon el resultado de su búsqueda:



Esta página de resultados le muestra información general concerniente a su búsqueda, esta vez mediante el identificador 2d1s, que corresponde a la molécula llamada Luciferasa.
Examine cuidadosamente la página de resultados y responda las preguntas a continuación:

¿En que fecha fue depositada esta estructura?
¿Cuál fue el método experimental mediante el que se obtuvo?
¿Qué significado cree usted que tienen los rótulos: "**Blue-Primary data**" y "**Red: Derived data**" que se encuentran en la parte superior izquierda?

En la parte superior derecha de esta página es posible visualizar esta estructura desde su navegador con cualquiera de los programas de visualización que allí se ofrecen (King, jmol, WbMol etc.).

Explore brevemente las diferentes opciones de visualización

En la parte superior de la página de resultados, encontrará una serie de pestañas que aportan mayor información acerca de la molécula:

Explore cada una de estas pestañas y explique, brevemente, qué información provee cada una de ellas.

Formato de archivo

Al determinar la estructura tridimensional de una proteína, obtenemos en realidad es información detallada de cada una de las coordenadas de sus componentes. Esta información se guarda en un archivo de texto, en un formato específico.

Presione el enlace "Download files", del menú de navegación a la izquierda.

Encontrará una serie de enlaces a archivos para descargar. De estos los más conocidos son **PDB** y **mmCIF**.

Presione el enlace "**PDB File**" y guardelo en su computador.

Este archivo es un archivo de texto, solamente que con extensión .PDB y puede ser abierto con cualquier editor de texto (Block de notas o Wordpad en sistemas MS windows o kate, kwrite, vim o Gedit en GNU/Linux).

Abra el archivo y examinelo cuidadosamente. Preste especial atención a las líneas que comienzan con la palabra ATOM. ¿Que información proveen?

Como podrá notarlo este archivo contiene bastante información y entenderla, por lo menos globalmente, resulta importante.

Para mayor información acerca de este formato, revise la siguiente guía explicativa que ofrece el sitio web de PDB:

http://www.wwpdb.org/documentation/format23/v2.3.html

Siga el enlace del menú izquierdo: "**structural analysis -> Geometry -> Molprobity Ramachandran plot"**. Esto generará un archivo descargable en formato PDF. Descarguelo y visualícelo. ¿Qué información provee este tipo de gráfico?

Uniprot

Visite la siguiente dirección: http://www.pir.uniprot.org/

Esa acción le llevará al sitio web de UNIPROT:

Uniprot es, en realidad, la reunión de varias bases de datos de proteínas. y se encuentra dividida en tres grandes secciones: Uniparc, UniprotKb y Uniref.

Haciendo uso de la sección "**About Uniprot -> Background**" explique en qué consiste cada una de estas divisiones.

Como podrá notarlo, en la parte superior derecha del sitio web de UNIPROT se encuentra una casilla de búsqueda.

Haga una búsqueda de la molécula de luciferasa (Luciferase en inglés). Presione el botón con una flecha a la derecha de esta casilla.

Después de uno segundos aparecerá la página de resultados:

Text Search Result (UniProtKB)

HELP

Search | Any Field | Restart | luciferase | + -- Add input box

928 entries found.

50 per page | **Page 1** > | Jump to Page | Display Options | Save Options | TABLE | FASTA | FLATFILE | XML

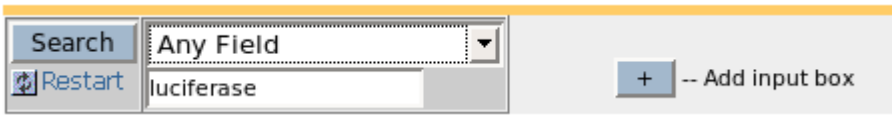
You may check one sequence and do BLAST or multiple sequences and do Multiple Alignment

ID/Accession	Protein Name	Length	Organism Name	Taxon Group	Gene Name	UniRef90/50	Matched Fields
LUCI_LUCCR / P13129	Luciferin 4-monooxygenase	548	Luciola cruciata	Euk/Animal		UniRef90_P13129 UniRef50_P00552	Protein Name=>Luciferase
LUCI_LUCLA / Q01158	Luciferin 4-monooxygenase	548	Luciola lateralis	Euk/Animal		UniRef90_P13129 UniRef50_P00552	Protein Name=>Luciferase
LUCI_LUCMI / Q26304	Luciferin 4-monooxygenase	548	Luciola mingrelica	Euk/Animal		UniRef90_Q26304 UniRef50_P00552	Protein Name=>Luciferase
LUCI_PHOPE / Q27757	Luciferin 4-monooxygenase	545	Photuris pennsylvanica	Euk/Animal		UniRef90_Q27757 UniRef50_P00552	Protein Name=>Luciferase
LUCI_PHOPY / P08659	Luciferin 4-monooxygenase	550	Photinus pyralis	Euk/Animal		UniRef90_P00552 UniRef50_P00552	Protein Name=>Luciferase
LUCI_RENRE / P27652	Renilla-luciferin 2-monooxygenase	311	Renilla reniformis	Euk/Animal		UniRef90_P27652 UniRef50_P27652	Protein Name=>luciferase
LUCI_VARHI / P17554	Luciferin 2-monooxygenase precursor	555	Vargula hilgendorfii	Euk/Animal		UniRef90_P17554 UniRef50_P17554	Protein Name=>luciferase
LUXA1_PHOLE / P09140	Alkanal monooxygenase	354	Photobacterium leiognathi	Bac/Gamma-proteo	luxA	UniRef90_P09140 UniRef50_P23146	Protein Name=>luciferase

¿Cuantas entradas arroja esta búsqueda?

La búsqueda que acaba de realizar es muy poco restringida, y de hecho lo que ha pasado es que el sistema de búsqueda de UNIPROT ha buscado el término “Luciferase” en cualquier campo de la base de datos y nos muestra los registros correspondientes.

Muchas veces esto no es lo que en realidad queremos, y necesitamos restringir nuestra búsqueda un poco más. Para esto podemos usar el formulario de búsqueda de esta página de resultados que se encuentra en la parte superior izquierda:

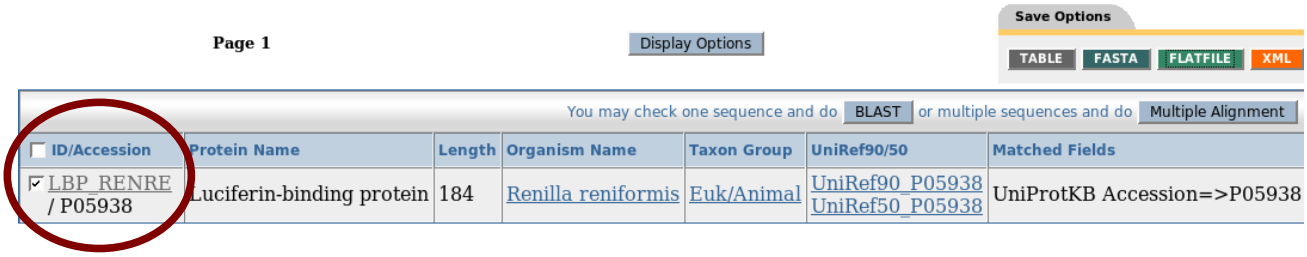


Realice la misma búsqueda, pero esta vez restringiendo las coincidencias al campo: “**Paper title”**, disponible en el menú desplegable del formulario. ¿Cuantos registros encontró esta vez?

Experimente con las diferentes opciones de filtro que ofrece este formulario de búsqueda.

Ahora realizaremos una búsqueda no tan abierta, sino a partir de un identificador ya conocido (de manera similar a como lo hicimos en PDB).

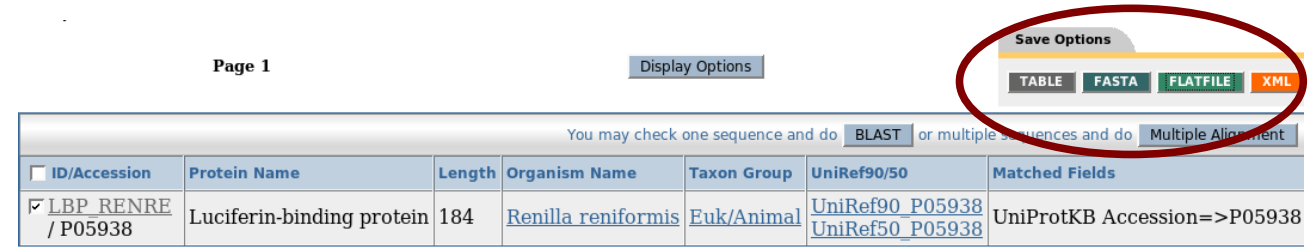
Realice la búsqueda de la siguiente molécula: P05938. Seleccione la entrada, haciendo click en la casilla de selección.



Descargue el archivo en formato "**flatfile**" correspondiente a esta entrada haciendo uso de la casilla de opciones de descarga:

Este es también un archivo de texto, puede abrirlo y examinarlo con su editor de textos preferido.

Experimente con los diversos formatos de descarga, ¿qué diferencias encuentra entre ellos?



Una vez descargue el archivo de resultados siga el enlace: "**LBP_RENRE"**.

En este momento debe encontrarse con la página de la entrada correspondiente a la molécula: "Luciferin-binding protein", la cual consta de varias secciones. Revise estas secciones cuidadosamente y responda las siguientes preguntas:

¿Cuál es la función de esta molécula?

¿Cuál es su nombre?

¿A que organismo corresponde?

¿Cuál es su peso molecular?

Esta guía es solamente un acercamiento inicial a PDB y UNIPROT, y no es posible detallar acá cada una de las posibilidades que estas bases de datos ofrecen. Por lo tanto, es importante que dedique tiempo extra y se familiarice más con ellas.

Guía elaborada por Andrés M. Pinzón V., del **Centro de Bioinformática** del Instituto de Biotecnología en la Universidad Nacional de Colombia y del **Laboratorio de Micología y Fitopatología** de la Universidad de los Andes, y está distribuida bajo licencia:

Bogotá Colombia – Febrero de 2007.

Cualquier sugerencia o inquietud dirigirla a:

ampinzonv@unal.edu.co ó andrespinzon@gmail.com