

Introducción a Bases de Datos (BD) Biológicas

Curso de Bioinformática

Luis E. Garreta U

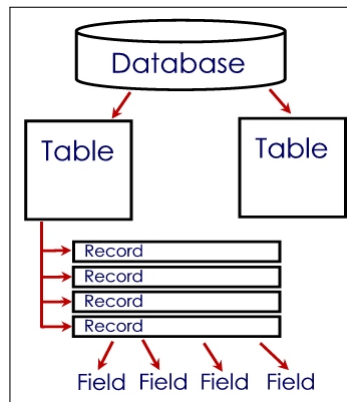
Pontificia Universidad Javeriana – Cali
Facultad de Ingeniería - Carrera de Biología

6 de agosto de 2018

Conceptos de BD

Qué es una Bases de Datos ?

- ▶ Una BD es una colección estructurada de información.
- ▶ La información o datos están organizados en **Tablas**.
- ▶ Las unidades básicas de una tabla son llamadas **registros**.
- ▶ Cada registro consiste de varios **campos**.
- ▶ Cada campo representa una información de ese registro.
- ▶ Cada registro tiene un identificador único.



Ejemplo: Base de Datos en forma de tabla

- ▶ Una BD se puede ver como una tabla donde:
 - ▶ Las filas corresponden a los registros
 - ▶ Las columnas corresponden a los campos

Campos

Field Record	Name	Length	Sequence	Enzyme
QA001	MTGA	243	MYQWI...	yes
QA002	Ribosomal protein L9	267	MAAPV...	no
QA003	Flagellin	374	GSSIL...	no
QA004	GDPMH	157	MFLRQ...	yes

Registros

↑

Identificador

- ▶ Número de Acceso (**Accession Number**): Son los identificadores únicos de cada registro de la BD

Organización de los datos en BD

- ▶ Archivos de texto planos
- ▶ BDs relacionales

Archivos de texto Planos

- Originalmente, todas las bases de datos utilizaban un formato de archivo de texto plano, sin formato.
- No contiene instrucciones ocultas para las computadoras.
- No se puede realizar búsquedas inteligentes por campos o por registros.

```

LOCUS      AAL93223              348 aa          linear   VRT 02-OCT-2003
DEFINITION NADH dehydrogenase subunit 2 [Ictalurus punctatus].
ACCESSION  AAL93223
VERSION    AAL93223.1  GI:19702261
DBSOURCE   accession AF482987.1
KEYWORDS   .
SOURCE     mitochondrion Ictalurus punctatus (channel catfish)
ORGANISM   Ictalurus punctatus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Siluriformes;
            Ictaluridae; Ictalurus.
REFERENCE  1 (residues 1 to 348)
AUTHORS    Waldbieser, G.C., Bilodeau, A.L. and Nonneman, D.J.
TITLE      Complete sequence and characterization of the channel catfish
            mitochondrial genome
COMMENT    Method: conceptual translation supplied by author.
ORIGIN
1 aspyvitill ssigigtalt fasshllaw ngleintlai lpmaeqhhp raveattkyf
61 laaaaaati lfastinawt tgewniycls hpaatilita slakvglep vhwappvmaq
121 glttttgllm atwqklapfa liiqapfth pllittlgll svfiggwgl nqtqlkila
181 yssiahlgwm iivtqykpql tvlvtityi mtsatfltfk laattkintl amswakvpti
241 tamaelalis lgglppttqf apkwlllqel taqgpltat amtisallsl yfyrlrcyam
301 titispntnn ssapwrlqnt qataplatm itallilpit plaqtltm
//
  
```

```

>hMLH1_wild-Type_Reference
GGAGGGACGAAGAGACCCAGCAACCCACAGAGTTGAGAAATTTGACTGGC
ATTCAAGCTGTCCAATCAATAGCTGCCGCTGAAGGGTGGGGCTGGATGGC
GTAAGCTACAGCTGAAGGAAGAAGCTGAGCACGAGGCACTGAGGTG
>HCT116_hMLH1_A01 <ID:1071327>
GGAGGGATGAAGAGATTAGTAATTTATAGAGTTGAGAAATTTGACTGGT
ATTTAAGTTGTTTAGTTAATAGTTGTTGTTGAAGGGTGGGGTGGATGGT
GTAAGTTATAGTTGAAGGACGAATGTGAGTATGAGGTATTGAGGTG
>HCT116_hMLH1_D01 <ID:1071363>
GGAGGGATGAAGAGATTAGTAATTTATAGAGTTGAGAAATTTGATTGGT
ATTCAAGTTGTTAATTAATAGTTGTTGTTGAAGGGTGGGGTGGATGGT
GTAAGTTATAGTTGAAGGAAGATGTGAGTATGAGGTATTGAGGTG
  
```

BDs Relacionales

- ▶ Contienen tanto los registros de datos como también las conexiones visibles e invisibles entre ellos.
- ▶ Esta organización facilitar la ejecución de las **consultas complejas**.

Table for **movie**

ID	Title	Year	Director
movie001	The man who shot Liberty Valance	1962	per001
movie003	The Grapes of Wrath	1940	per001
movie003	Pulp Fiction	1994	per002
movie004	The matrix	1999	per003
movie005	Cidade de Deus	2002	per004

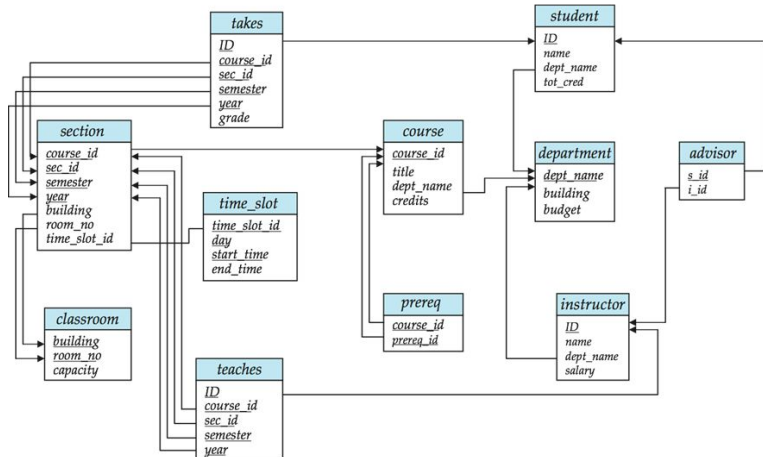
relation

Table for **person**

ID	Name	Surname	Origin
per001	John	Ford	USA
per002	Quentin	Tarantino	USA
per003		Wachowski	USA
per004	Fernando	Mirelles	Brazil

Esquema de una BD Relacional


- Muchas tablas, con muchos registros, con muchos campos y muchas conexiones



BDs Biológicas

Luis E. Garreta U

Crecimiento de las BDs Biológicas



GenBank Overview

[PubMed](#)
[Entrez](#)
[BLAST](#)
[OMIM](#)
[Books](#)
[Taxonomy](#)
[Structure](#)

Search for

NCBI Home

NCBI Site Map

GenBank Submissions Handbook

Submit to GenBank

Submit an update

Search GenBank

GenBank and RefSeq: a comparison

BLAST

What is GenBank?

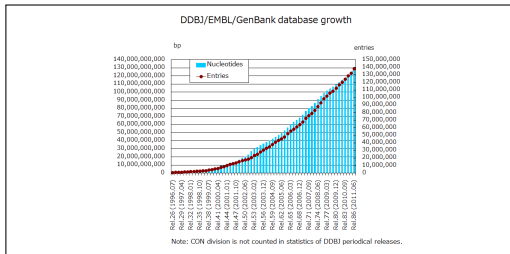
GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2011 Jan 39(Database issue):D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site.

A new release is made every two months. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank [record](#) may be viewed for a *Saccharomyces cerevisiae* gene.

Existen aproximadamente 286,730,369,256 registros de secuencias en las divisiones tradicionales del GenBank (2011).



Crecimiento por Divisiones del GenBank

Medida en pares de bases de nucleótidos (bp)

Table 1. Growth of GenBank Divisions (nucleotide base-pairs)

Division	Description	Release 173 (8/2009)	Release 179 (8/2010)	Increase (%)
TSA	Transcriptome shotgun data	39 829 979	398 676 845	900.9
ENV	Environmental samples	1 091 072 890	1 723 286 428	57.9
PAT	Patented sequences	5 592 927 651	8 519 294 473	52.3
BCT	Bacteria	4 107 328 206	5 333 010 385	29.8
VRL	Viruses	779 481 462	970 125 245	24.5
PHG	Phages	36 100 172	43 456 808	20.4
MAM	Other mammals	576 977 646	679 274 390	17.7
INV	Invertebrates	1 734 996 371	2 036 240 836	17.4
WGS	WGS data	148 165 117 763	169 253 846 128	14.2
GSS	Genome survey sequences	16 738 219 857	18 442 479 673	10.2
PLN	Plants	3 695 552 256	4 038 424 961	9.3
SYN	Synthetic	131 361 806	142 548 355	8.5
VRT	Other vertebrates	2 366 300 257	2 533 789 261	7.1
EST	ESTs	34 522 977 161	36 803 930 321	6.6
HTC	High-throughput cDNA	636 472 189	659 355 057	3.6
PRI	Primates	5 751 413 009	5 943 029 356	3.3
ROD	Rodents	4 206 718 960	4 298 354 944	2.2
HTG	High-throughput genomic	23 895 733 886	24 276 862 305	1.6
UNA	Unannotated	119 348	120 289	0.8
STS	Sequence tagged sites	629 573 650	634 263 196	0.7
TOTAL	All GenBank sequences	254 698 274 519	286 730 369 256	12.6

Registros de un BD Biológica

- Ejemplo de un registro en formato GenBank

```
LOCUS      AAL93223                348 aa          linear   VRT 02-OCT-2003
DEFINITION NADH dehydrogenase subunit 2 [Ictalurus punctatus].
ACCESSION  AAL93223
VERSION    AAL93223.1  GI:19702261
DBSOURCE   accession AF482987.1
KEYWORDS   .
SOURCE     mitochondrion Ictalurus punctatus (channel catfish)
ORGANISM   Ictalurus punctatus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Siluriformes;
            Ictaluridae; Ictalurus.
REFERENCE  1 (residues 1 to 348)
AUTHORS    Waldbieser, G.C., Bilodeau, A.L. and Nonneman, D.J.
TITLE      Complete sequence and characterization of the channel catfish
            mitochondrial genome
COMMENT     Method: conceptual translation supplied by author.
ORIGIN
1  mspyviti ll  sslglgtalt fmsshwlaw ngleintlai iplnaqhnhp raveattkyf
61  laqaaaaati lfastinawt tgewniycis hpaatilitm alaalkvglap vhfwnppvmaq
121  gltlttglin atwqklapfa liiqmapfth plllttlgll svfiggwggll ngtqlrkile
181  yssiahlgum iivtqykpql tvlvityii mtsatfltfk laatktintl amswakvpti
241  tamaaalais lggllppltgf mpkwlllqel tmqglpitat antlsallsl yfyrlrcyan
301  titispntnn ssapwrlqnt qataplaltm intllllplt plaqtltm

//
```

Campos Esenciales en un registro de una BD biológica

- ▶ La secuencia
- ▶ El número de acceso o *Accession Number* (AC)
- ▶ Datos taxonómicos
- ▶ Referencias
- ▶ Palabras clave (keywords)
- ▶ Documentación/ Anotaciones/Curación (curation)
 - ▶ Grupos de investigación (sumisión directa)
 - ▶ información literatura suplementaria
 - ▶ Intitutos de secuenciación
 - ▶ Patentes

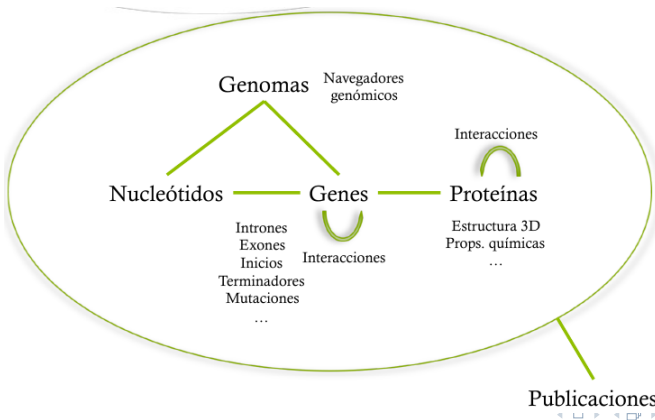
El Formato de los registros necesita ser consistente dentro de la BD

Un entrada en la BD de proteínas SwissProt en formato **FASTA**:

```
>sp|P01588|EPO_HUMAN Erythropoietin OS=Homo sapiens OX=9606 GN=EPO  
PE=1 SV=1  
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAEENITTGCAE  
HCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQ  
PWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPDAASAAPLRTITADTFRKLFRV  
YSNFLRGKCLKLYTGEACRTGDR
```

Importancia de las BDs Biológicas?

- ▶ El propósito de estas BD va más allá de simplemente almacenar datos de forma organizada.
- ▶ Se busca también permitir una recuperación de datos **inteligente**.



Recuperación de Información a través de Consultas o Queries

- ▶ Un **query** o **consulta** es el método para recuperar información de la BD.
- ▶ Especialmente las **consultas sobre los campos**, debido a la organización de los registros con base a **campos**.

The screenshot shows the NCBI Structure database search results for the query 'villin'. The interface includes a search bar with 'villin' entered, a 'Structure' dropdown menu, and buttons for 'Create alert' and 'Advanced'. Below the search bar, there are options for 'Summary' (20 per page) and 'Sort by Default order'. The search results are displayed as a list of items, with the first three items shown. Each item includes a checkbox, a thumbnail image of a protein structure, and a title link. The first item is 'Crystal structure of calcium-free human gelsolin[Actin binding Protein]', the second is 'ATP bound gelsolin[CONTRACTILE PROTEIN, STRUCTURAL PROTEIN]', and the third is 'The Crystal Structure Of Calcium-Free Equine Plasma Gelsolin[Contractile Protein]'. Each item also includes taxonomic information, protein IDs, and links to 'View in iCn3D', 'Similar Structures', 'PubMed', 'Proteins', and 'Conserved Domains'.


NCBI Resources How To


Structure Structure villin Create alert Advanced


Summary 20 per page Sort by Default order Send to:

Search results

Items: 1 to 20 of 116 << First < Prev Page 1 of 6 Next > Last >>

1.  [Crystal structure of calcium-free human gelsolin\[Actin binding Protein\]](#)
Taxonomy: Homo sapiens
Proteins: 1 modified: 2017-11-20
MMDB ID: 77149 PDB ID: 3FFN
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#)

2.  [ATP bound gelsolin\[CONTRACTILE PROTEIN, STRUCTURAL PROTEIN\]](#)
Taxonomy: Equus caballus
Proteins: 2 Chemicals: 2 modified: 2018-01-08
MMDB ID: 38727 PDB ID: 2FGH
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#) [PubChem Compound](#)

3.  [The Crystal Structure Of Calcium-Free Equine Plasma Gelsolin\[Contractile Protein\]](#)
Taxonomy: Equus caballus
Proteins: 2 modified: 2012-11-27
MMDB ID: 11164 PDB ID: 1D0N
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#)

Tipos de BDs Biológicas

Tipos de BDs Biológicas

- ▶ BDs Primarias
- ▶ BDs Secundarias
- ▶ BDs Especializadas

BDs Primarias

- ▶ Contienen datos biológicos originales
- ▶ Secuencias crudas o datos estructurales sometidos por la comunidad científica
- ▶ Estas son:
 - ▶ **GenBank**: mantenida por el NCBI (National Center for Biotechnology Information)
 - ▶ **EMBL**: mantenida por el EBI (European Bioinformatics Institute)
 - ▶ **DDBJ**: DNA Database of Japan
 - ▶ **PDB**: Protein Data Bank mantenida por el RCSB (Research Collaboratory for Structural Bioinformatics)

BDs Secundarias

- ▶ Las BD secundarias contienen información procesada computacional o manualmente por un experto, a partir de información original de las bases de datos primarias
- ▶ Las BD de secuencias traducidas de proteínas que contienen anotaciones funcionales pertenecen a esta categoría
- ▶ Algunos ejemplos son:
 - ▶ **UniProt**: un recurso completo, de alta calidad y de libre acceso de secuencias proteicas e información funcional
 - ▶ **PIR** (Protein Information Resources) que es sucesor del Atlas of Protein Sequence and Structure

BDs Especializadas

- ▶ Las BD especializadas son aquellas dedicadas un interés de investigación particular
- ▶ Por ejemplo:
 - ▶ Flybase: Una BD de genes y genomas de Drosófila
 - ▶ HIV sequence database,
 - ▶ Ribosomal Database Project
 - ▶ ...

BD GenBank:

<https://www.ncbi.nlm.nih.gov/genbank/>

The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar labeled 'GenBank' with a dropdown menu set to 'Nucleotide' and a 'Search' button. A horizontal menu below the search bar lists various data types: GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, and Other. The main content area is titled 'GenBank Overview' and includes a section 'What is GenBank?' which describes the database as a collection of publicly available DNA sequences. It mentions its affiliation with the International Nucleotide Sequence Database Collaboration (INSDC) and lists its partners: DNA DataBank of Japan (DDBJ), European Nucleotide Archive (ENA), and GenBank at NCBI. A paragraph follows about the release schedule (every two months) and where to find release notes and growth statistics. Another paragraph mentions an annotated sample record for *Saccharomyces cerevisiae*. To the right of the overview, there's a 'GenBank Resources' section with links to 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'. Below the overview, an 'Access to GenBank' section states that there are several ways to search and retrieve data, followed by a bulleted list of methods: searching with Entrez Nucleotide, aligning with BLAST, downloading with NCBI e-utils, and using the anonymous FTP server.

NCBI Resources How To

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan 41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

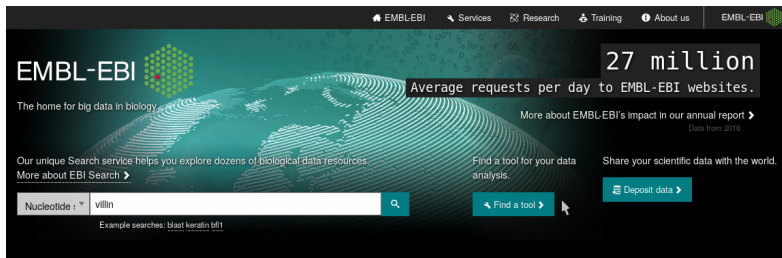
- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utils](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

BD EMBL-EBI:

<https://www.ebi.ac.uk/>



The screenshot shows the EMBL-EBI homepage with a dark teal background and a globe graphic. The navigation bar at the top includes links for EMBL-EBI, Services, Research, Training, and About us. The main header features the EMBL-EBI logo and the text "The home for big data in biology". A prominent statistic states "27 million Average requests per day to EMBL-EBI websites." Below this, there are three main sections: "Our unique Search service helps you explore dozens of biological data resources." with a link to "More about EBI Search", a search bar with "villin" entered and a dropdown menu for "Nucleotide", and "Example searches: blast kernal bll1"; "Find a tool for your data analysis." with a link to "Find a tool"; and "Share your scientific data with the world." with a link to "Deposit data".

We are EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact. ➤


Data resources

Explore our open data resources to enrich your research. Browse data, perform analyses or share your own results. ➤

Research

Find out about our research groups, postdoctoral schemes and PhD Programme ➤

DDBJ: <https://www.ddbj.nig.ac.jp>

 Services ▾[Login & Submit](#) [Contact](#) [Japanese](#)

DDBJ Center

DDBJ Center Web Sites ▾



DDBJ Nucleotide Sequence Submission System (NSSS) will be unavailable (Aug. 15(Wed)15:00 - Aug. 21(Tue)10:00)

DDBJ Center provides sharing and analysis services for data from life science researches and advances science.

Search & Analysis



Submissions



Downloads



SuperComputer



Statistics



Activities



Training



About Us



News from DDBJ Center

3 August 2018 | [Maintenance](#) | [DDBJ](#) | [BioProject](#) | [BioSample](#) | [DRA](#)

PDB:

RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ More ▾

RCSB
PDB
PROTEIN DATA BANK142602 Biological
Macromolecular Structures
Enabling Breakthroughs in
Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands

[Advanced Search](#) | [Browse by Annotations](#)

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

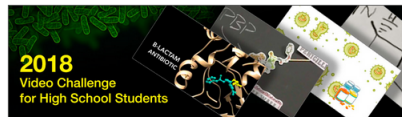
A Structural View of Biology

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

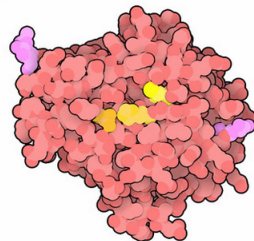
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Award-Winning Videos on Antibiotic Resistance



August Molecule of the Month



Legumain

Ejercicio: Realizar búsquedas en todas las BD primarias

Buscar los términos:

"early-onset breast cancer human brca2"

- *BRCA2* y BRCA2 son el gen y su proteína, respectivamente para el gen relacionado con el cancer de mama.

Como debería ser una "buena BD" biológica

- Completa, pero fácil de realizar búsquedas

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"
- ▶ Simple y fácil de entender su estructura

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"
- ▶ Simple y fácil de entender su estructura
- ▶ Con referencias cruzadas (*cross-referenced*)

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"
- ▶ Simple y fácil de entender su estructura
- ▶ Con referencias cruzadas (*cross-referenced*)
- ▶ Redundancia mínima

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"
- ▶ Simple y fácil de entender su estructura
- ▶ Con referencias cruzadas (*cross-referenced*)
- ▶ Redundancia mínima
- ▶ Fácil recuperación de los datos

Como debería ser una "buena BD" biológica

- ▶ Completa, pero fácil de realizar búsquedas
- ▶ Anotada, pero no "demasiado anotada"
- ▶ Simple y fácil de entender su estructura
- ▶ Con referencias cruzadas (*cross-referenced*)
- ▶ Redundancia mínima
- ▶ Fácil recuperación de los datos

Problemas con las BD Generales (Primarias)

BDs que se esfuerzan por una "amplitud enciclopédica" son demasiado grandes y se vuelven inmanejables:

- ▶ Propensas a ser bastante redundantes
- ▶ Secuencias inadecuadas:
 - ▶ Viejas
 - ▶ Parcialmente anotadas
 - ▶ Inconsistentes y anotaciones viejas
 - ▶ Secuencias con errores o de baja calidad
 - ▶ Contaminaciones
 - ▶ Secuencias anónimas