

Máster Universitario en Bioinformática

Generación y mantenimiento de datos ómicos

Curso académico 2024-25



Universidad
Internacional
de Valencia

Dr. Jordi Tronchoni
jordi.tronchoni@professor.universidadviu.com

23/04/2024

De:
 Planeta Formación y Universidades

Tema 4

Bases de datos y herramientas bioinformáticas

23/04/2024

Contenidos

Tema 1. Introducción a la bioinformática

- 1.1 Historia de la bioinformática
- 1.2 Bioética aplicada al análisis de datos

Tema 2. Principales flujos de trabajo en bioinformática

- 2.1 Genómica
- 2.2 Metagenómica y metataxonómica
- 2.3 Transcriptómica
- 2.4 Proteómica

Tema 3. Gestión de entornos y paquetes

- 3.1 Conda

Tema 4. Bases de datos y herramientas bioinformáticas

- 4.1 Principales bases de datos
- 4.2 Otros recursos online

Tema 5. Alineamiento de secuencias

- 5.1 Introducción al alineamiento de secuencias
- 5.2 Alineamientos Pairwise
- 5.3 Alineamientos Múltiples

Tema 6. Métodos de secuenciación

- 6.1 Primera generación de secuenciadores
- 6.2 Segunda generación de secuenciadores
- 6.3 Tercera generación de secuenciadores
- 6.4 Comparación de plataformas de secuenciación

Tema 7. Pre-procesado y calidad de secuencias

- 7.1 Calidad de secuencias
- 7.2 Pre-procesado de secuencias

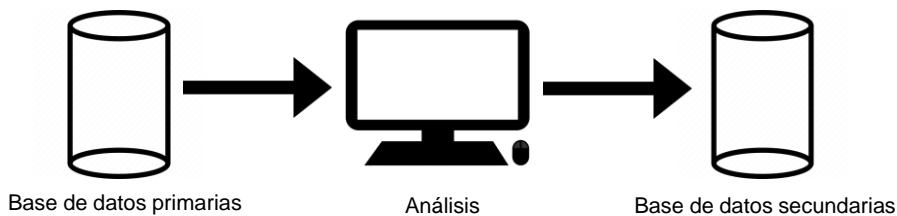
Bases de datos y herramientas bioinformáticas

Las secuencias generadas se alojan en bases de datos dentro de cada uno de los campos de la bioinformática

Su principal función es la de almacenar toda la información generada a través de las estrategias de secuenciación y ensamblaje para poder caracterizar la nueva información que se generará en un futuro

- Bases de datos primarias
- Bases de datos secundarias
- Bases de datos especializadas

Generalmente la información es ofrecida como texto plano



23/04/2024

Toda la información generada a través de la secuenciación y análisis bioinformático se almacena en bases de datos. Generalmente cada una de las técnicas bioinformáticas tiene su conjunto de bases de datos, permitiendo no solo almacenar la nueva información generada si no la comparación y caracterización de nueva información respecto a lo almacenado. Estas bases de datos se dividen en tres grupos: bases de datos primarias, secundarias y especializadas. Las bases de datos primarias almacenan lo generado mediante las diversas técnicas, mientras que las secundarias almacenan las información procesada desde las bases de datos primarias. Por su parte, las especializadas reúnen información sobre, por ejemplo, organismos modelos concretos.

Bases de datos y herramientas bioinformáticas

- Una de las principales herramientas en biología son las bases de datos
- Incluyen datos de secuencias de nucleótidos, proteínas, estructura de proteínas, ...etc
- Concebidas para la consulta humana, distintos programas capaces de leerlas
- Generales o específicas

Principales bases de datos de secuencias de nucleótidos:

-  **Enlace 1**
NCBI (EEUU)
<http://www.ncbi.nlm.nih.gov/>
-  **Enlace 2**
EMBL (Europa)
<http://www.ebi.ac.uk/embl/>
-  **Enlace 3**
DDBJ (Japón)
<http://www.ddbj.nig.ac.jp>

Algunas bases de datos de genomas de organismos concretos:

-  **Enlace 4**
Flybase (Drosophila)
<http://flybase.bio.indiana.edu/>
-  **Enlace 5**
SGD (Levadura)
<http://www.yeastgenome.org/>
-  **Enlace 6**
TAIR (Arabidopsis)
<http://www.arabidopsis.org/>

Principales bases de datos de proteínas:

-  **Enlace 8**
Uniprot
<http://www.pir.uniprot.org/>
-  **Enlace 9**
Swiss-prot
<http://expasy.org/sprot/>
-  **Enlace 10**
PDB
<http://www.rcsb.org/pdb/home/home.do>

23/04/2024

Dentro de las bases de datos primarias podemos encontrar bases de datos de nucleótidos, de proteínas y de estructuras. Las principales bases de datos primarias son GenBank, pertenecientes al grupo del NCBI; ENA, del grupo EBI; y DDBJ, "DNA Data Bank of Japan". La particularidad de estas bases de datos es que comparten información, sincronizándose entre ellas cada 24h. Por otra parte, la principal base de datos de proteínas es, a día de hoy, Uniprot. En esta base de datos tenemos información acerca de la secuencia de aminoácidos de todas las proteínas desde secuencias codificantes alojadas en las bases de datos de nucleótidos. La base de datos de estructura tridimensional de macromoléculas PDB almacena información relacionada con las coordenadas atómicas de las estructuras proteínas medidas por cristalografía de rayos X y RMN.

Bases de datos y herramientas bioinformáticas

- Una de las principales herramientas en biología son las bases de datos
- Incluyen datos de secuencias de nucleótidos, proteínas, estructura de proteínas, ...etc
- Concebidas para la consulta humana, distintos programas capaces de leerlas
- Generales o específicas

Bibliografía:



Enlace 13

Pubmed

<http://www.NCBI.nlm.nih.gov/sites/entrez>



Enlace 14

Scopus

<https://www.Scopus.com/>



Enlace 15

Web of Knowledge

<http://www.webofknowledge.com/>

Rutas metabólicas:



Enlace 16

KEGG

<http://www.genome.jp/kegg/>

Enfermedades genéticas humanas:



Enlace 17

OMIM

<http://www.NCBI.nlm.nih.gov/sites/entrez?db=omim>

Bases de datos y herramientas bioinformáticas

Principales bases de datos de secuencias de nucleótidos:



Enlace 1

NCBI (EEUU)

<http://www.NCBI.nlm.nih.gov/>



Enlace 2

EMBL (Europa)

<http://www.ebi.ac.uk/embl/>

- EMBL
 - EBI
- NCBI




4.1 Principales bases de datos

EMBL

NCBI

En general podemos referirnos a ellas como Recursos de datos o más exactamente, recursos de datos y herramientas.



EMBL → EBI → *data resources*

En general podemos referirnos a ellas como Recursos de datos o más exactamente, recursos de datos y herramientas.

Bases de datos y herramientas bioinformáticas

- Laboratorio Europeo de Biología Molecular (EMBL)

Institución de investigación en biología molecular financiada por 18 países europeos que fue creada en 1974. Tiene laboratorios en Heidelberg (Alemania), Hamburgo (Alemania), Grenoble (Francia) y Hinxton (Reino Unido); un programa de investigación en Monterotondo (Italia), y **una subsede en Barcelona** (España), desde el 23 de diciembre de 2016.

Funciones del EMBL:

- ✓ la realización de investigación básica en biología molecular;
- ✓ la formación y entrenamiento de científicos, estudiantes y visitantes;
- ✓ el patrocinio de programas activos de ciencia en la sociedad.

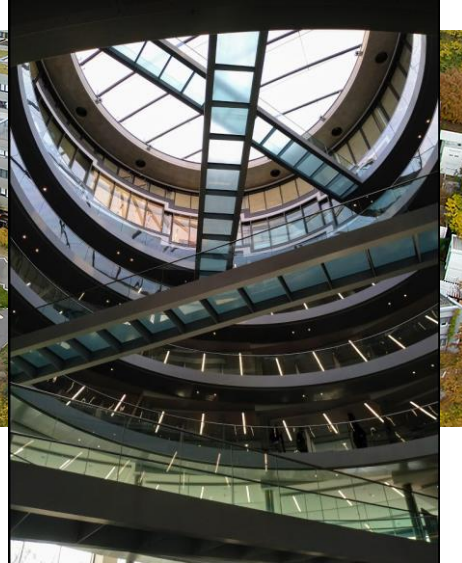
23/04/2024

EMBL – Heidelberg - GER



23/04/2024

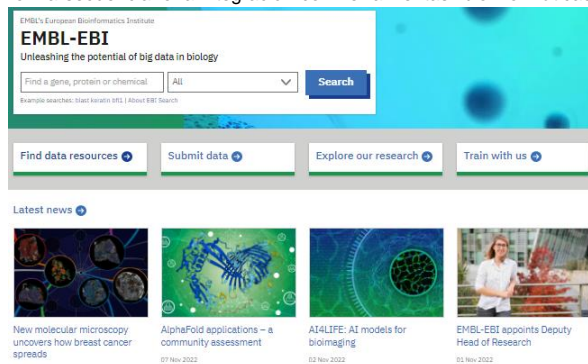
EMBL – Heidelberg - GER



Bases de datos y herramientas bioinformáticas

- Instituto Europeo de Bioinformática (EBI) (EMBL's European Bioinformatics Institute)
 - ✓ Líder a escala mundial en desarrollo y mantenimiento de recursos bioinformáticos.
 - ✓ Gama más amplia del mundo de recursos de datos moleculares disponibles de forma gratuita y actualizada.
 - ✓ Servicios web que permiten consultar sus bases de datos biológicas de manera programada, y posibilitan el desarrollo de análisis de forma secuencial o la integración con herramientas bioinformáticas de análisis propios y de otras instituciones.

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



23/04/2024

EBI - Hinxton - UK



23/04/2024

<https://www.ebi.ac.uk/services/data-resources-and-tools>

EMBL-EBI data resources and tools

EMBL's European Bioinformatics Institute maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Search

23/04/2024

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical

Example searches: blast keratin bf11 | About EBI Search

Find data resources →

Latest news →

Search

All

Science search

- Genomes & metagenomes
- Nucleotide sequences
- Protein sequences
- Small molecules
- Gene expression
- Gene-Disease Associations
- Diseases
- Molecular interactions
- Reactions & pathways
- Protein families
- Literature
- Samples & ontologies

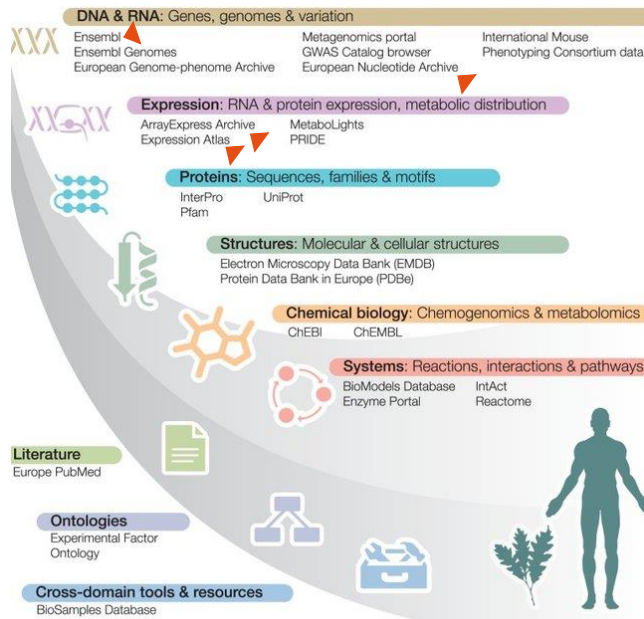
Search web content

- EMBL-EBI People
- EMBL-EBI web

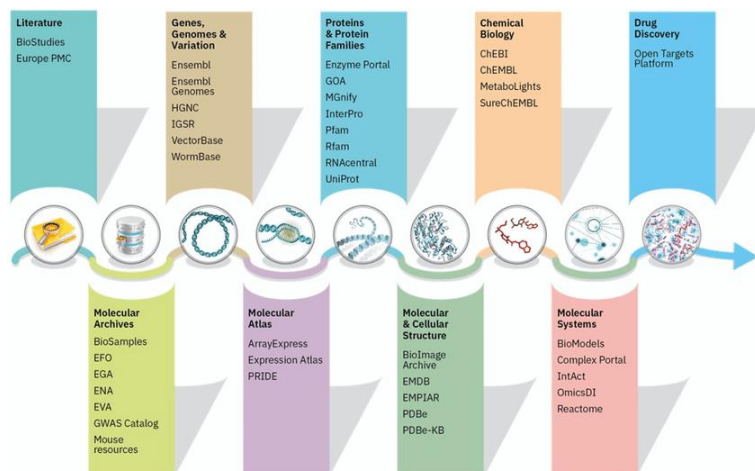
Explore our r

23/04/2024

Data resources

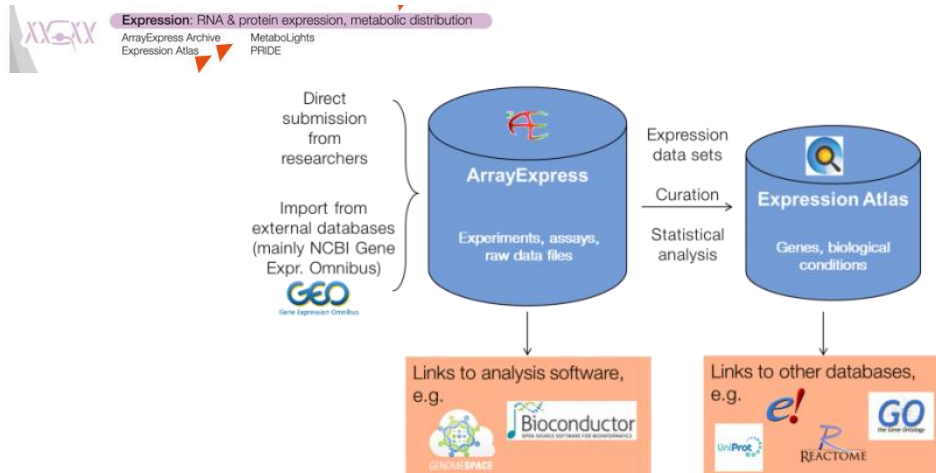


Data resources



23/04/2024

Bases de datos y herramientas bioinformáticas



23/04/2024

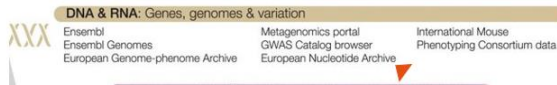
En realidad tendemos a llamar de forma generalista a todo bases de datos, y esto genera cierta confusión. Además las bases de datos tienen numerosos enlaces y a veces es difícil saber donde estas o porque un link te ha llevado donde te ha llevado. Aumenta la confusión que los datos estén duplicados en numerosas bases de datos...

Por ejemplo, ArrayExpress o Expression Atlas, son dos bases de datos. Dos bases de datos independientes, pero altamente conectadas. Mientras que ArrayExpress contiene Experimentos, o archivos en bruto, Expression Atlas contiene los resultados de análisis de ArrayExpress. En ambas bases de datos encontraremos múltiples links que nos llevaran fuera de la base de datos a herramientas de análisis de los datasets donde nos encontramos o a otras bases de datos ligadas al dataset que estábamos analizando.

ArrayExpress y Expression Atlas son dos bases de datos que se encuentran entre los recursos de datos del EBI, el Instituto Europeo de Bioinformática, que pertenece (cuelga de) al EMBL, el Laboratorio Europeo de Biología Molecular.

Debido a ello, la denominación, recursos de datos y herramientas bioinformáticas es más acertado.

Bases de datos y herramientas bioinformáticas



- **European Nucleotide Archive (ENA)**, del Laboratorio Europeo de Biología Molecular (EMBL) e Instituto Europeo de Bioinformática (EBI).
<https://www.ebi.ac.uk/ena/browser/home>

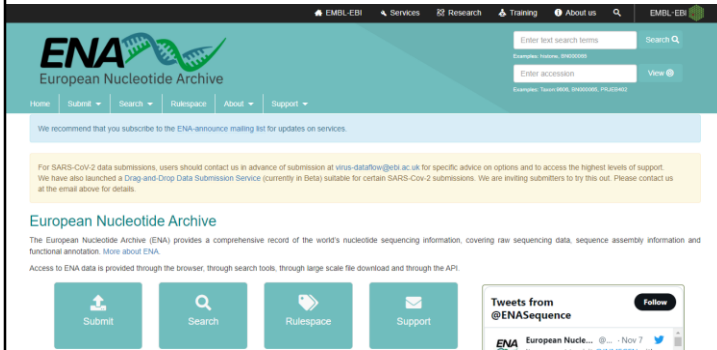
ENA forma parte -actualmente es su mayor contribuyente: **>925 millones de secuencias de ADN registradas-** de la **International Nucleotide Sequence Database Collaboration (INSDC)**, que también incluye el **GenBank del NCBI** -con >761 millones de secuencias- y el **DNA Data Bank of Japan (DDBJ)** -que contiene ~76 millones de secuencias.

23/04/2024

La base de datos ENA es la base de nucleótidos europea. Presenta una interfaz diferente a la del NCBI pero su contenido es similar. Presenta una particularidad, y es que almacena secuencias EST procedentes de transcriptomas

Bases de datos y herramientas bioinformáticas

European Nucleotide Archive (ENA)



- Secuencias de ADN y ARN anotadas.
- Información complementaria:
 - ✓ detalles del ensamblaje de secuencias
 - ✓ Procedimientos experimentales
 - ✓ Otros metadatos relacionados con los proyectos de secuenciación (muestra, configuración experimental, configuración de la máquina, mapeo, ensamblaje, etc.)

23/04/2024

European Nucleotide Archive (ENA)

Enlace de interés

<https://www.ebi.ac.uk/ena/submit/data-formats>

En esta página web se pueden ver los datos de **ENA** organizados en una serie de clases distintas divididas en tres niveles (**reads**, **assembly**, **annotation**), cada una de las cuales tiene sus propios formatos y estándares.

23/04/2024

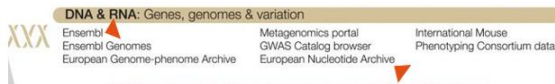
Organización de los datos

• How to Explore an ENA Project

Data Domain	Description	Example Records
Projects/Studies	Contains information on a biological research project. This holds all the data generated as part of this research	PRJEB1787 (ERP001736)
Samples	Represents biological samples collected and sequenced in real life	SAMEA2620084 (ERS488919)
Reads (Runs/Experiments)	Hold raw read files and sequencing methods	ERR1701760 ERX1772048
Analyses	Hold results files of analyses performed on sequencing data and analysis methods	ERZ1195979
Contig set	Hold contig sets generated as part of a genome or transcriptome assembly.	CABH0Y010000000.1
Assemblies	Represents an entire genome assembly and holds any contig sets or sequence records generated as part of the assembly	GCA_000001405.28
Assembled/Annotated Sequences (*)	Any sequence records from coding or non-coding regions to full assembled chromosomes	CM000667.2
Taxon	The sequenced organism or metagenome of a sample	Taxon:9606
Sample Checklist	The checklist of metadata that the sample was registered with	ERC000013

* Assembled and annotated sequence records fall into different data classes. Read more about the different classes of sequences [here](#).

Bases de datos y herramientas bioinformáticas



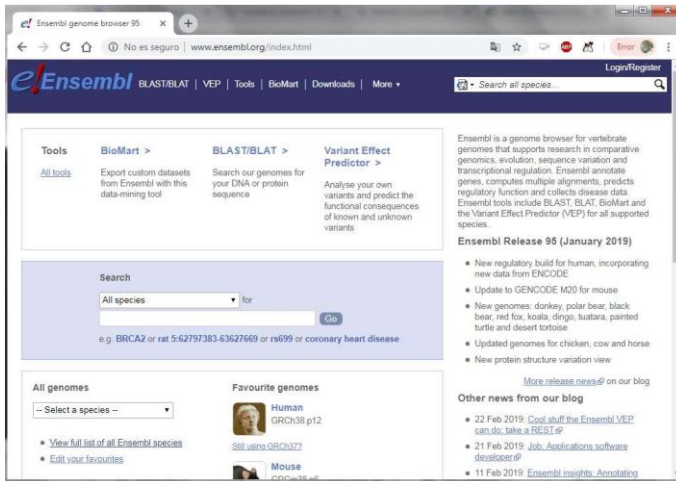
- **Ensembl**, del Laboratorio Europeo de Biología Molecular (EMBL) e Instituto Europeo de Bioinformática (EBI).

<http://www.ensembl.org/index.html>

Base de datos de genomas completos de organismos vertebrados a través de la cual es posible realizar investigación en **genómica comparativa, evolución, variación de secuencia y regulación transcripcional**. Ensembl **anota genes**, permite efectuar **alineaciones múltiples, predecir la función reguladora de un gen** y **recopilar datos de enfermedades**.

Bases de datos y herramientas bioinformáticas

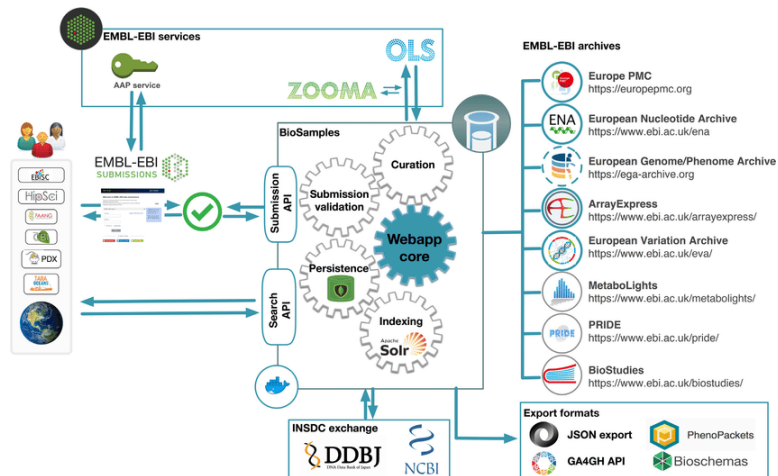
Ensembl



- Incluye programas bioinformáticos, como, por ejemplo, **BLAST/BLAT**, **BioMart** y **Variant Effect Predictor (VEP)**, cuyas descripciones y usos aparecen detallados en los enlaces correspondientes.

23/04/2024

Bases de datos y herramientas bioinformáticas



23/04/2024

El proceso de subida de datos a las bases de datos es complejo para intentar mantener un orden y una clasificación adecuada. Son muchas las solicitudes de subida de datos que se producen ya que para poder publicar un trabajo científico, un “paper” los datos deben de encontrarse en un repositorio público.

Data submission

[EMBL-EBI home](#) > [Services](#) > [Data submission](#)

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

[DNA/RNA sequence](#)

[Expression data](#)

[Protein data](#)

[Structures](#)

[Systems](#)

[Chemical biology](#)

[Ontologies](#)

[Images](#)

[Multi-omics or other cross-domain study](#)

[Other biological research data](#)

Why submit data to an archive?

- Submission of primary data and derived information to public data repositories is an essential step in the scientific process.
- Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems.
- Serving as a complement to the literature publication process and supporting early data sharing, the EMBL-EBI offers a number of submission services appropriate for different types and scales of data.

Need help?

If you need help with your data submission, please [contact support](#).

All EMBL-EBI data repositories

[Array Express](#): functional genomics data

[BioImage Archive](#): bioimaging data

[BioModels](#): computational models

[BioSamples](#): reference sample data

[BioStudies](#): biological research data

[ChEBI](#): chemical entities

[DGVa](#): structural genetic variation data

[EFO](#): experimental variables

[EGA](#): human data that requires controlled access

[EMPIAR](#): raw image data

[ENA](#): nucleotide sequence data

[EVA](#): genetic variation data

[GO](#): Gene ontology annotations

[GWAS Catalog](#): Genome-wide association study data

[IntAct](#): molecular interactions

[IntEnz](#): enzyme nomenclature

[MetaboLights](#): metabolomics data

[Metagenomics](#): raw sequence data & associated meta-data

23/04/2024

Para ello el EBI cuenta con una herramienta para “facilitar” la subida de los datos de forma correcta, esto es bien etiquetados (metadatos) al repositorio correcto.

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



[EMBL-EBI home](#) > [Services](#) > [Data submission](#)

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

[DNA/RNA sequence](#)

[Expression data](#)

[Protein data](#)

[Structures](#)

[Systems](#)

[Chemical biology](#)

[Ontologies](#)

[Images](#)

[Multi-omics or other cross-domain study](#)

[Other biological research data](#)

Why submit data to an archive?

- Submission of primary data and derived information to public data repositories is an essential step in the scientific process.
- Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems.
- Serving as a complement to the literature publication process and supporting early data sharing, the EMBL-EBI offers a number of submission services appropriate for different types and scales of data.

Need help?

If you need help with your data submission, please [contact support](#).

All EMBL-EBI data repositories

[Array Express](#): functional genomics data

[BioImage Archive](#): bioimaging data

[BioModels](#): computational models

[BioSamples](#): reference sample data

[BioStudies](#): biological research data

[ChEBI](#): chemical entities

[DGVa](#): structural genetic variation data

[EFO](#): experimental variables

[EGA](#): human data that requires controlled access

[EMPIAR](#): raw image data

[ENA](#): nucleotide sequence data

[EVA](#): genetic variation data

[GO](#): Gene ontology annotations

[GWAS Catalog](#): Genome-wide association study data

[IntAct](#): molecular interactions

[IntEnz](#): enzyme nomenclature

[MetaboLights](#): metabolomics data

[Metagenomics](#): raw sequence data & associated meta-data

23/04/2024

A pesar de lo que indican, la realidad es otra, es complejo y bastante tedioso, razón por la que mayoritariamente, no encontramos los datos subidos correctamente o etiquetados correctamente. Es muy recomendable no dejar este paso para el final, hacerlo correctamente lleva tiempo. Es interesante revisar toda la documentación antes de abordar el deposito de los datos.



NCBI

Bases de datos y herramientas bioinformáticas

- Centro Nacional para la Información Biotecnológica de los Estados Unidos (NCBI).

Es parte de la Biblioteca Nacional de Medicina de los Estados Unidos (National Library of Medicine), una rama de los Institutos Nacionales de Salud (National Institutes of Health, NIH). Está localizado en Bethesda, Maryland, y fue fundado el 4 de noviembre de 1988 con **la misión de ser una importante fuente de información y desarrollo en biología molecular.**

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

Bases de datos y herramientas bioinformáticas

- Centro Nacional para la Información Biotecnológica de los Estados Unidos (NCBI).

NCBI Resources How To Sign In to NCBI

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Biosays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
Novel coronavirus complete genome from the Wuhan outbreak now available in GenBank
13 Jan 2020
The immediate sequenced genomes
The new NHMS system launch is just around the corner
10 Jan 2020
Mark your calendars! As announced last month, a new NIH Manuscript
Recent improvements to the genome Assembly resource
09 Jan 2020
We're constantly making improvements to the NCBI coronavirus Assembly resource
[More...](#)

23/04/2024

Bases de datos y herramientas bioinformáticas

- Centro Nacional para la Información Biotecnológica de los Estados Unidos (NCBI).

El **NCBI** ofrece un número variado de recursos científicos, p. ej.:

GenBank (almacena y constantemente actualiza información referente a secuencias genómicas)

PubMed (contiene resúmenes y citas de artículos científicos referentes a biomedicina, biotecnología, bioquímica, genética y genómica)

OMIM (contiene una recopilación de enfermedades genéticas humanas)

dbSNP (base de datos de polimorfismos de nucleótidos simples)

RefSeq (incluye la anotación precisa de genes conocidos y permite la predicción de nuevos genes basados en la evidencia de transcripción disponible)

BLAST (alineamiento múltiple de secuencias). <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Bases de datos y herramientas bioinformáticas

- Centro Nacional para la Información Biotecnológica de los Estados Unidos (NCBI).

Todas las bases de datos del NCBI están disponibles en **línea de manera gratuita** y son accesibles usando el buscador **Entrez** (barra de búsqueda localizada en la parte superior central de cada una de las bases de datos).

Bases de datos y herramientas bioinformáticas

Genbank

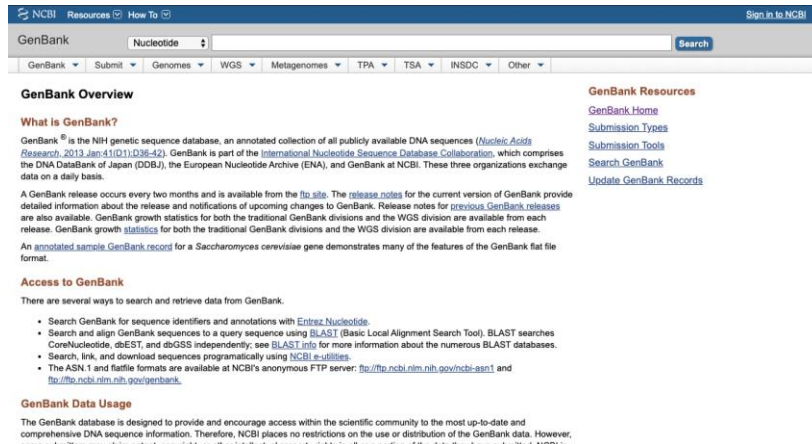
- Creada en 1982 por Walter Goad y el Laboratorio Nacional de Los Álamos
- **>761 millones de secuencias de ADN de más de 300.000 especies** registradas
- es parte también de la **International Nucleotide Sequence Database Collaboration (INSDC)**.
- crece a un ritmo exponencial (se duplica aprox cada 18 meses).
- obtiene sus datos principalmente a través de los **registros de laboratorios individuales** y de **lotes de proyectos de secuenciación a gran escala** que incluyen proyectos de WGS y proyectos de muestreo ambiental.
- la mayoría de los registros se realizan utilizando los programas **BankIt**, **Tbl2asn**, **Submission Portal** y **Sequin**, y el personal de GenBank asigna números de acceso a los datos registrados.

23/04/2024

Bases de datos y herramientas bioinformáticas

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 with a dropdown menu set to 'Nucleotide' and a 'Search' button. A secondary navigation bar includes links for '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?' with a paragraph explaining its role as the NIH genetic sequence database. It also mentions the 'GenBank release' cycle and provides links to 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'. A section titled 'Access to GenBank' lists several ways to search and retrieve data, including using Entrez Nucleotide, BLAST, and various file formats. Finally, a 'GenBank Data Usage' section states that the database is designed to provide and encourage access within the scientific community.

NCBI Resources How To Sign in to NCBI

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 Acid Research](#), 2013 Jan 41(1):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](#).
- 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-utilities](#).
- 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 Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, users are asked to acknowledge the source of the data when they publish their work.

GenBank Resources

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

23/04/2024

Bases de datos y herramientas bioinformáticas

Genbank

Ejemplo de un registro de datos en GenBank (**U49845**, relativo al gen **TCP1-beta** de ***Saccharomyces cerevisiae***) con el formato *GenBank Flat Format* (GFF):

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

NCBI Resources How To

GenBank Nucleotide


GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

Sample GenBank Record

This page presents an annotated sample GenBank record (accession number **U49845**) in its *GenBank Flat File* format. You can see the corresponding [live record for U49845](#), and see [examples of other records](#) that show a range of biological features.

LOCUS	SCU49845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.				
ACCESSION	U49845				
VERSION	U49845.1 GI:1293613				
KEYWORDS	.				
SOURCE	Saccharomyces cerevisiae (baker's yeast)				
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.				
REFERENCE	1 (bases 1 to 5028)				
AUTHORS	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.				
TITLE	Cloning and sequence of REV7, a gene whose function is required for				

23/04/2024



Uniprot; PDB; KEGG; SGD...

Bases de datos biológicas

Bases de datos primarias

UniProt (Universal Protein Resource) - EBI + Swiss Institute of Bioinformatics (SIB) + Georgetown University.
Aúna Swiss-Prot + TrEMBL

- Swiss-Prot : Proteínas curadas manualmente
- TrEMBL: Automáticamente

UniProt Knowledgebase (UniProtKB) – Secuencias proteínas y sus anotaciones

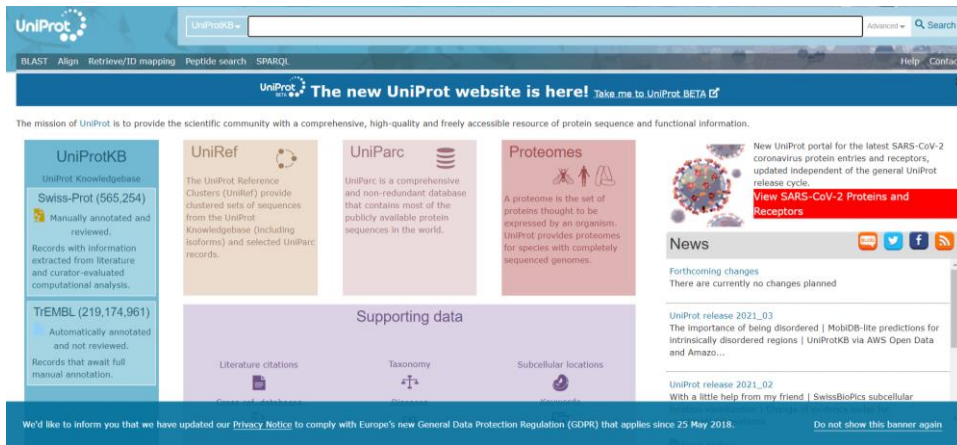
UniProt Reference Clusters Database (UniRef)

23/04/2024

UniProt es la principal base de datos de proteínas la cual aúna información de otras bases de datos. Es una combinación de distintos institutos, como el EBI, el instituto suizo y la universidad de Georgetown. Esta base de datos aúna información de Swiss-Prot y TrEMBL. Estas dos bases de datos se diferencian principalmente porque Swiss-Prot revisa manualmente toda la información anotada en la misma, mientras que en TrEMBL se analiza de forma automática.

Bases de datos biológicas

Bases de datos primarias



23/04/2024

Esta es la imagen que presenta su página de inicio.

Bases de datos biológicas

Bases de datos primarias

The screenshot shows the UniProtKB website interface. At the top, there's a navigation bar with links like BLAST, Align, Retrieve/ID mapping, Peptide search, and SPARQL. A banner announces 'The new UniProt website is here!'. The main heading is 'UniProtKB - I3WFC7 (I3WFC7_9BETA)'. On the left, a 'Display' sidebar lists various views: Entry (selected), Publications, Feature viewer, and Feature table. The main content area shows protein details: 'Protein: Capsid scaffolding protein', 'Gene: N/A', 'Organism: Macacine betaherpesvirus 3 (Rhesus cytomegalovirus)', and 'Status: Unreviewed - Annotation score: 4.00 - Protein inferred from homology'. The 'Function' section describes the protein's role in capsid assembly and DNA packaging.

23/04/2024

Podemos buscar el nombre de una proteína y acceder a su ficha.

Bases de datos biológicas

Bases de datos secundarias

Prositate (expasy)

Una base de datos proteicas clasificada por motivos proteicos, relacionados con función

Un motivo deriva de alineamientos múltiples y se guarda la consenso

Expasy

PROSITE

[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)



Database of protein domains, families and functional sites



SARS-CoV-2 relevant PROSITE motifs

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / [References](#) / [Commercial users](#)].
PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionality and/or structurally critical amino acids [More...].

Release 2021_03 of 02-Jun-2021 contains 1891 documentation entries, 1311 patterns, 1321 profiles and 1333 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] [Examples](#)

(Enter UniProtKB accession or identifiers or PDB identifiers)

Other tools

- PRATT - allows to interactively generate conserved patterns from a series of unaligned proteins.

23/04/2024

Entre las bases de datos secundarias encontramos aquellas que derivan información desde la secuencia de aminoácidos. Prosite es una de ellas, y guarda información de motivos proteicos, los cuales se relacionan con la función de las proteínas



PROSITE documentation PDOC00022

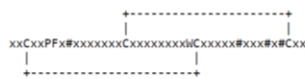
Fibronectin type-II collagen-binding domain signature and profile

Description Technical section References Copyright Miscellaneous

Description

Fibronectin is a plasma protein that binds cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. The major part of the sequence of fibronectin consists of the repetition of three types of domains, which are called type I, II, and III [1]. Type II domain (FN2) is approximately 40 residues long, contains four conserved cysteines involved in disulfide bonds and is part of the collagen-binding region of fibronectin [2]. In fibronectin the minimal collagen binding region is formed by one FN1 and two FN2 domains. This suggests that the collagen-binding sites spans multiple modules.

A schematic representation of the position of the invariant residues and the topology of the disulfide bonds in FN2 domain is shown below.



'C': conserved cysteine involved in a disulfide bond.
'#': large hydrophobic residue.

The 3D-structure of the FN2 domain has been determined (see <PDB:2FN2>) [3]. The structure consists of two double-stranded anti-parallel β sheets oriented approximately perpendicular to each other and two intrachain loops connecting the two β sheets and the


23/04/2024

Aquí podemos observar un dominio de la proteína fibronectina tipo 2. Podemos leer información sobre la proteína y una representación del motivo. Este motivo está descrito y hace referencia a funciones concretas donde la proteína está involucrada.

Bases de datos biológicas

Bases de datos secundarias

KEGG



KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND DISEASE DRUG

Select prefix: map Organism Enter keywords: Go Help

[New pathway maps | Update history]

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

- 1. Metabolism**
Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Organismal Systems**
- 6. Human Diseases**
- 7. Drug Development**

KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

Pathway Identifiers

Each pathway map is identified by the combination of 2-4 letter prefix code and 5 digit number (see KEGG Identifier). The prefix has the following meaning:

- map manually drawn reference pathway
- ko reference pathway highlighting KOs
- ec reference metabolic pathway highlighting EC numbers
- m reference metabolic pathway highlighting reactions
- <org> organism-specific pathway generated by converting KOs to gene identifiers

and the numbers starting with the following:

- 011 global map (lines linked to KOs)
- 012 overview map (lines linked to KOs)
- 010 chemical structure map (no KO expansion)

23/04/2024

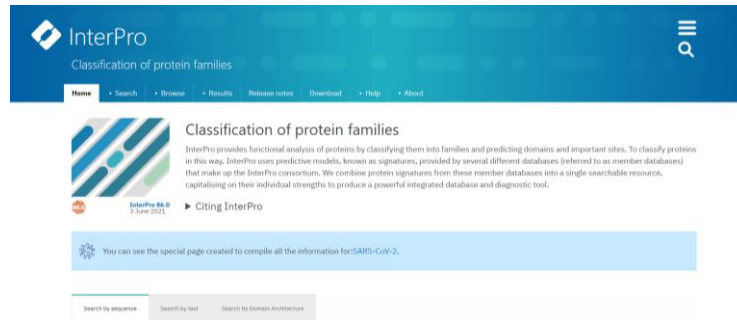
La base de datos KEGG contiene información sobre rutas metabólicas y las macromoléculas que están involucradas. De esta forma podemos predecir, en función de la información existente en la base de datos, cuales son las funciones relacionadas de nuestras proteínas.

Bases de datos biológicas

Bases de datos secundarias

Interpro

Metabase de datos que aúna otras bases de datos como Swiss-Prot, Pfam o TIGRFAMs



23/04/2024

Interpro por su parte aparece como una asociación de bases de datos proteicas y bases de datos funcionales, como Swiss-Prot, Pfam y TIGRFAM, es un ejemplo de metabase de datos.

Bases de datos biológicas

Bases de datos secundarias

Interpro

InterPro Classification of protein families

Home • Search • Browse • Results • Release notes • Download • Help • About

Home / Browse / By Entry / InterPro / IPR000003 / Overview

F Retinoid X receptor/HNF4[®] IPR000003

InterPro entry

Overview Proteins 3K Taxonomy 1K Proteomes 451 Structures 126 AlphaFold 25 Pathways 34

Short name: Retinoid-X_rcpt/HNF4

Overlapping homologous superfamilies

- Zinc finger, NHR/GATA-type (IPR013090)
- Nuclear hormone receptor-like domain superfamily (IPR035500)

Family relationships

- Nuclear hormone receptor (IPR001723)
- Retinoid X receptor/HNF4 (IPR000003)

Description

Steroid or nuclear hormone receptors (NRs) constitute an important superfamily of transcription regulators that are involved in widely diverse physiological functions, including control of embryonic development, cell differentiation and homeostasis. Members of the superfamily include the steroid hormone receptors and receptors for thyroid hormone, retinoids, vitamin D, and a variety of other ligands (1). The prototype function is described

Add your annotation

Contributing Member Database Entry

PRINTS

PRINTS: PR00545

23/04/2024

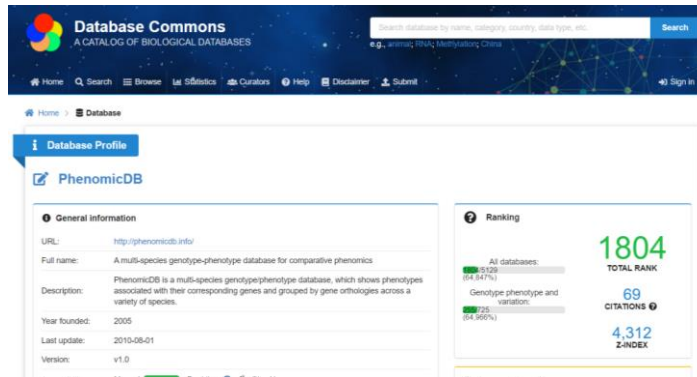
Aquí tenemos un ejemplo del receptor X retinoide.

Bases de datos biológicas

Bases de datos secundarias

PhenomicDB

Base de datos que contiene datos de humano y organismos modelos



23/04/2024

Existen las bases de datos con información sobre el fenotipo. En este ejemplo tenemos la base de datos PhenomicDB, con información de humano y organismos modelos

Bases de datos biológicas

Bases de datos secundarias

Estructuras moleculares

PDB (Protein data bank)

Determinación experimental de la cristalografía de macromoléculas. Generalmente hay proteínas, pero también hay ADN, ARN y complejos híbridos



23/04/2024

También existen las bases de datos de estructuras, siendo la principal PDB (Protein data bank). Esta base de datos reúne la información relacionada con estructura de proteínas, ADN o ARN mediante cristalografía de rayos X o RMN.

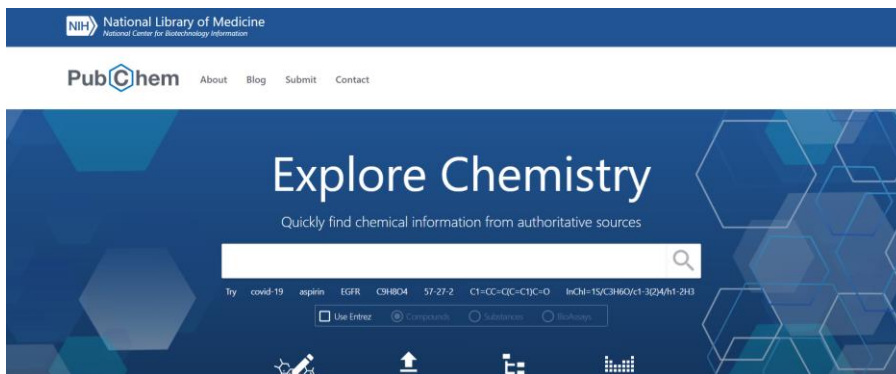
Bases de datos biológicas

Bases de datos secundarias

Estructuras moleculares

PubChem

Compuestos químicos e información sobre su actividad



23/04/2024

También tenemos la base de datos PubChem, con información de compuestos químicos.

Bases de datos biológicas

Bases de datos especializadas

Base de datos de The Saccharomyces Genome Database (SGD)

SGD Saccharomyces GENOME DATABASE

Analysis Sequence Function Literature Community

Search SGD, SGD, SGD, SGD, SGD

About SGD

The Saccharomyces Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast *Saccharomyces cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.

Explore SGD

Meetings

Fungal Pathogen Genomics
May 07 to May 12, 2023 -
Wellcome Genome Campus, Hinxton, Cambridge, UK

CSHL Yeast Genetics & Genomics course
July 26 to August 15, 2023 -
Cold Spring Harbor Laboratory, New York

TAGC2024 The Allied Genetics Conference
March 09 to March 10, 2024 -
National Harbor | Washington DC Metro Area

New & Noteworthy

Regulation of yeast inducible promoters by SAGA and TFIIID - November 15, 2022
Gene transcription – the elaborate process that our cells use to read genetic information stored in DNA – was long thought to be turned on only when certain regulatory factors traveled to specific DNA sequences. In a new study published in *Genes & Development*, Mittal et al., 2022, discovered that a subset of genes has their transcription [...] [Read More](#)

2D-RNA structures from RNAcentral - October 20, 2022
SGD has updated our RNA pages to add secondary structures provided by RNAcentral and generated by R3DIT. Thumbnails and links to RNAcentral

Tweets from @yeastgenome

13 SGD Project Reviewed
D. Alan Drummond Art
@dadrummondart · Nov 15
Inside a cell, my outway model of a dividing cell with all its delicate machinery, mother producing daughter in late anaphase. Expect the mitochondria... Available at [thepw.mn/1rge](#) if you'd like your own!

23/04/2024

Existen repositorios o bases de datos especializadas en organismos modelo. Estas bases de datos reúnen información de organismos concretos como por ejemplo *Saccharomyces*. La base de datos SGD guarda toda la información relevante sobre este organismo, como genomas, genes, proteínas, fenotipos...

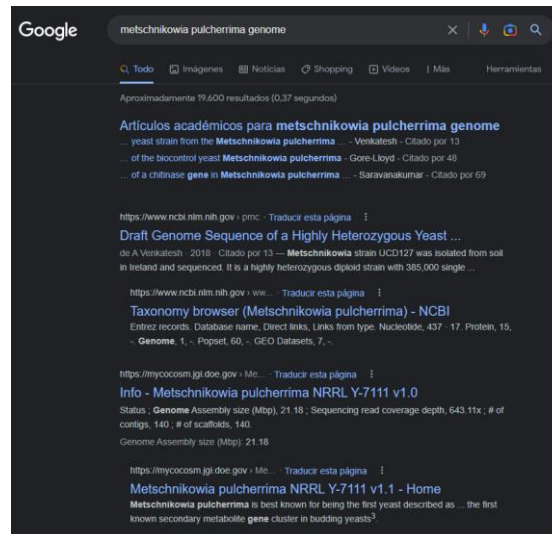
Bases de datos biológicas

Principales problemas de las bases de datos

- Secuencias con errores o con secuencias de vectores de clonación, adaptadores, etc...
- Información mal anotada, campos incompletos
- Sesgos sobre microorganismos muy estudiados, redundancia

23/04/2024

Las bases de datos por lo general tienen una serie de errores o problemas sistemáticos. Estos problemas derivan en su mayoría de la falta de criterio unificado a la hora de añadir nueva información a las mismas y el tipo de estudios que se realizan. Podemos encontrar secuencias con errores, los cuales deberán curarse manualmente al detectarse, o secuencias mal anotadas como vectores de secuenciación anotados como plásmidos de secuencias bacterianas. También podemos tener problema en la información que se muestran en las fichas para cada organismo o proteína, ya que a pesar de tener la información más relevante, por ser obligatoria, podemos estar perdiendo la información de, por ejemplo, la fuente desde donde se aisló la muestra. Por último podemos observar problemas en la sobrerrepresentación de diversos organismos, como los organismos modelos o bacterias (como E.coli), lo que puede generar un problema en la caracterización de nuevas secuencias en las bases de datos.



23/04/2024

Es importante conocer como se estructura una base de datos y donde esta la información, pero a la hora de buscar *algo* sigue siendo más sencillo pasar por un buscador.

Por ejemplo, digamos que me estoy embarcando en un nuevo proyecto que va a implicar la secuenciación de una cepa de levadura de *Metschnikowia pulcherrima*. Ciertamente podría ir a la base de datos de genomas del EBI o del NCBI y buscar por organismo, y probablemente sería muy recomendable hacerlo así... pero en realidad, solemos empezar por escribir en el buscador:
Metschnikowia pulcherrima genome

La razón es que muy probablemente vaya a localizar esos genomas depositados en las principales bases de datos pero... puede que exista algún repositorio más pequeño, alguna universidad o proyecto que tenga los datos públicos o disponibles. En este caso, mycocosm, e indagando vemos que esta actualizado con un genoma con PacBio en el 2020.



4.2 Otros recursos online

Herramientas online

Software para ensamblaje y edición de fragmentos de ADN

gatk

User Guide Tool Index Blog Forum Events Download GATK Sign in

Genome Analysis Toolkit

Variant Discovery in High-Throughput Sequencing Data

Sequencing → Reads → gatk best practices → Variants

Developed in the Data Sciences Platform at the Broad Institute, the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping. Its powerful processing engine and high performance computing features make it capable of taking on projects of any size. [Learn more](#)

Find answers to your questions. Stay up to date on the latest topics. Ask questions and help others.

- Getting Started**
Best practices, tutorials, and other info to get you started
- Technical Documentation**
Algorithms, glossary, and other detailed resources
- Announcements**
Blog and events
- Tool Index**
Purpose, usage and options for each tool
- Forum**
Ask our team for help and report issues
- GATK Showcase on Terra**
Check out these fully configured workspaces
- Download latest version of GATK**
The GATK package download includes all released GATK tools. Please see Getting Started for a quick start primer on using GATK, and see the Version History for release notes and highlights of differences between versions.
- Run on Cloud**
- Run on HPC**

23/04/2024

Herramientas online

PCA transcriptómica

PCAGO

About Analyse Help Impressum

PCAGO helps you analyzing your RNA-Seq read counts with principal component analysis (PCA). We also included other helpful features like **read count normalization**, downloading annotations and GO terms for your genes and a tool to find a gene variance cut-off for PCA.

Start analysis!
or learn how to use it!

PROCESS YOUR DATA
Upload your read count data or choose one of our example data sets. If your data is not normalized, you

FILTER YOUR GENES
You can filter the genes by scaffold, biotype or GO terms. Just upload an annotation or use the

CLUSTER AND PCA
Create interactive PCA and hierarchical clustering plots of your sample and use the animation feature to analyze

<https://pcago.bioinf.uni-jena.de/>

23/04/2024

Herramientas online

U.S. National Library of Medicine | NCBI | National Center for Biotechnology Information | tronchoni | My NCBI | Sign Out

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

A new version IgBLAST (1.17) is here.
We've added a new field "V frame shift" to the IgBLAST output to indicate if there is an internal frame shift in the normal V gene translation frame.
Thu, 14 Jan 2021 12:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

BLAST Genomes

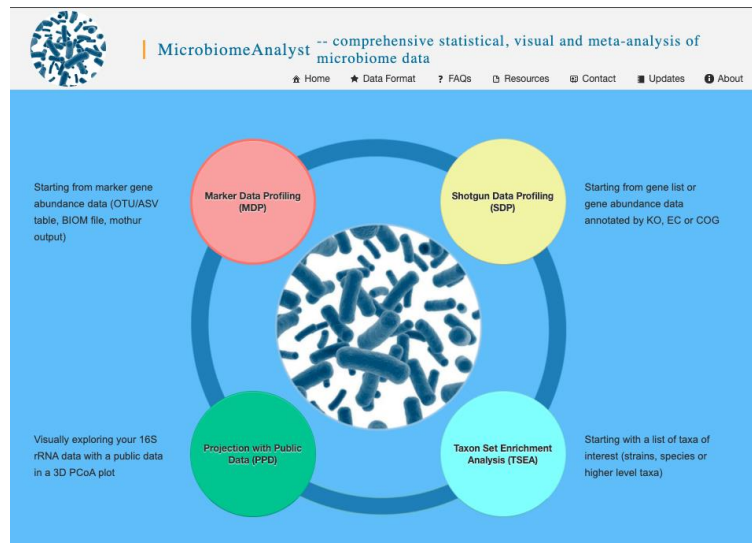
Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes

23/04/2024

Herramientas online

Software para metataxonomía



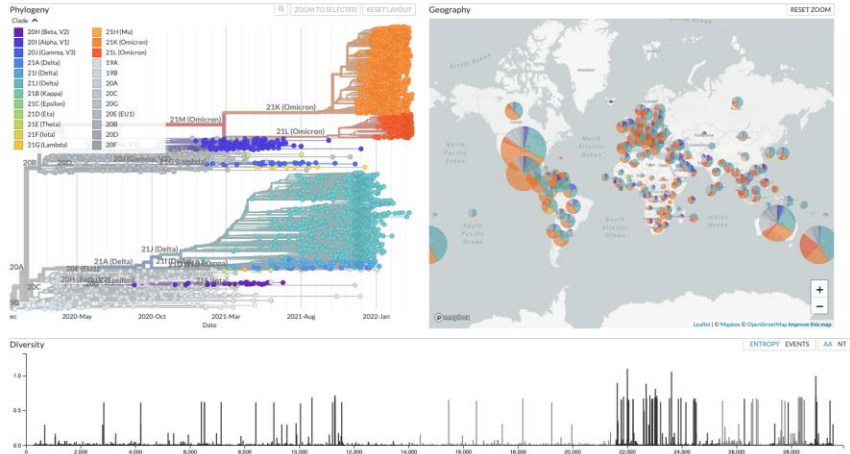
Herramientas online

Epidemiología Genómica de SARS-CoV-2 NextStrain

Genomic epidemiology of SARS-CoV-2 with global subsampling

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from GISAID.

Showing 3002 of 3002 genomes sampled between Dec 2019 and Mar 2022.



23/04/2024

Existen otro tipo de recursos online

https://www.embl.org/	EMBL
http://www.ebi.ac.uk/	EBI
https://www.ncbi.nlm.nih.gov/	
https://www.ddbj.nig.ac.jp/index-e.html	DDBJ
https://www.ebi.ac.uk/training/online/course-list	ebi course
https://www.ncbi.nlm.nih.gov/books/NBK21101/	ncbi books
https://www.iscb.org/cms_addon/online_courses/index.php	Iscb course
http://biochem218.stanford.edu/index.html	Stanford course
https://www.huber.embl.de/msmb/introduction.html	Modern statistics book
http://seqanswers.com/	Bioinformatics forums
https://www.biostars.org/	& communities

23/04/2024

Es importante conocer los dos grandes foros de bioinformática, seqanswers y biostars. Generalmente cuando tenemos un problema de instalación o hay alguna herramienta que no funciona adecuadamente, ya le ha ocurrido a alguien y es aconsejable revisar estos foros para poder encontrar la solución.

También podemos encontrar interesantes discusiones sobre la utilización de los distintos parámetros de distintas herramientas y conocer nuevas formas de realizar distintos procesos.

¡Gracias!



Universidad
Internacional
de Valencia

universidadviu.com

De:
🌐 Planeta Formación y Universidades