Biological Databases

How the biological information is stored?

Biological Databases

1. Introduction to BDs
2. kind and classification of DBs
3. Main BioDBs
4. DBs Structures
5. Other DBs

¿What is biological database?

A **database** is an organized collection of data, so that it can be easily accessed and managed.
You can organize data into tables, rows, columns, and index it to make it easier to find relevant information.

- A biological DB is a collection of biological data that include different dataset with different characteristics which are related among each other and stored under an specific purpose and logical structure
- Database = Structure + Data
- Example: a Library

History of BioDBs

- **What came first: Gene or protein?**

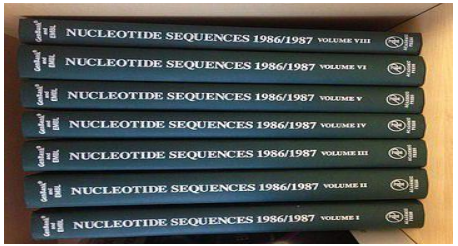
- **BD without internet?**

Every year, the first issue of the journal Nucleic Acids Research (NAR) is dedicated to databases: it publishes articles that describe the creation of new databases and the innovations that have occurred in existing ones, as well as contains an exhaustive list of all existing databases and their URLs. Many of these databases are hosted on the websites of government or private centers that have created a uniform graphical environment that brings together a large number of databases.

Bases de Datos Núcleotidicas

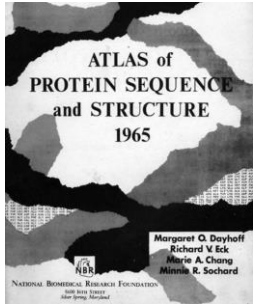
- In 1982 the EMBL (Heidelberg-Germany) begins to store DNA sequences. Followed by GeneBank US and DDBJ.
- In 1988 the three groups decide to create the INSDC. International Nucleotide Sequence Database Collaboration and unify format.
- In 2003 NCBI starts the RefSeq project.Objective is to reduce the enormous redundancy of the INSDC.
- 2002-2003 the development of the Genome-Browsers begins. Map-Viewer (NCBI), ENSEMBL (EBI-Sanger Center)

GeneBank 86/87



Protein Data Base

- 1960 Dayhoff et. to the. they put together all the known prot sequences in the Protein Information Resource (PIR). This derived in the “Atlas of Protein Sequence and Structure” obtained by traditional methods that included annotations and was kept in printed form until 1972 and later it was transferred to magnetic tape.
- In the 80s Swiss-Prot begins to work, when “Aimos Bairoch” converts the PIR Atlas to a format similar to that of EMBL.
- In 1986 Bairoch distributed SwissProt in the forerunner of the internet, at that time there were 3900 streams.
- In about 2000 SwPr was complemented with TrEMBL, with automatic annotations.
- In 2004 Creates UniProt gathering and processing the information of SwPr, TrEMBL and PIR



Information in the genomic era

Since the human genome project and other species the flow of information is extremely huge

In order to use the information has to be correctly stored and structured

The access to the information must be

- easy
- fast
- flexible
- friendly
- reliable
- robust

Not all the biological DBs accomplish these features

¿What are BioDBs for

Which biological information to search?

- Bibliography
- organism and associated information
- genes and derived elements (mRNA, CDS, proteins, enzymes, etc)
- genomes
- genes annotations
- genome annotations
- Biomolecules
- structures
- Taxonomy
- complex biological information

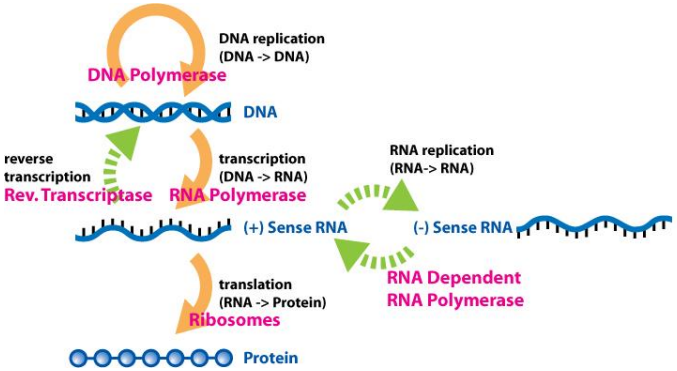
complex information deserves further structure

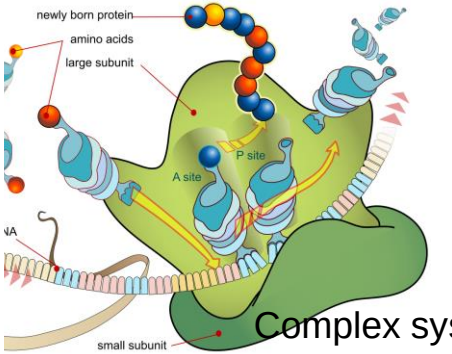
- Gene regulation
- Interactions:
    - gene-gene
    - gene-protein
    - protein-protein
    - protein-biomolecules
- metabolic pathways
- Reactions
- Environmental features

of course, to search biological information!!

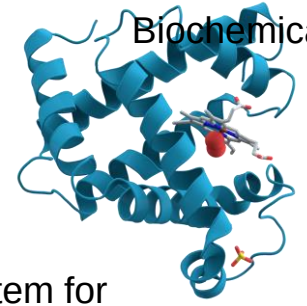
Scheme for non biologist

Central dogma of molecular biology

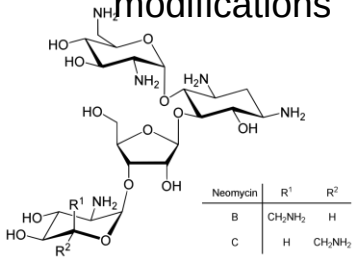




Protein structure
Biochemical interactions

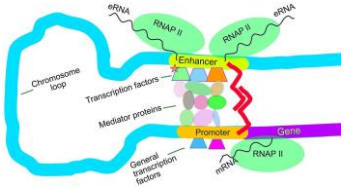


Post-traductional modifications

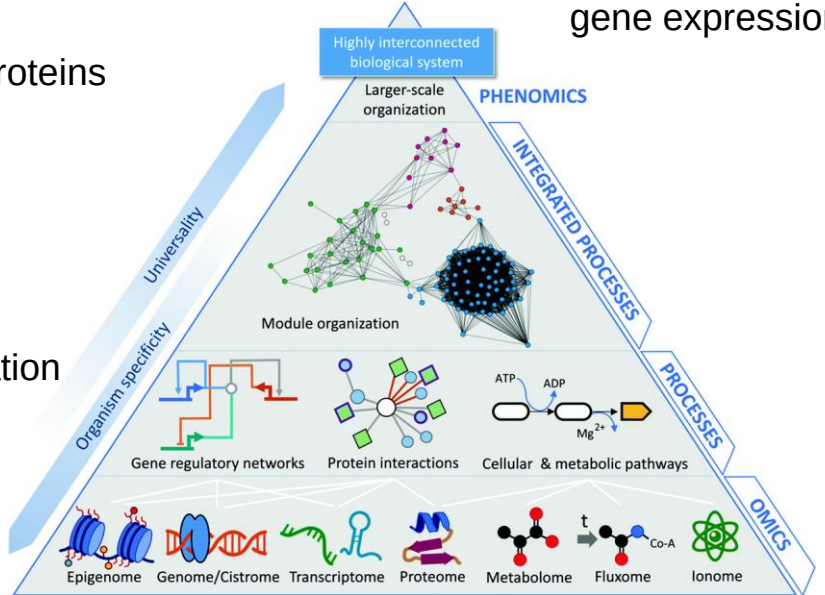


Complex system for gene expression

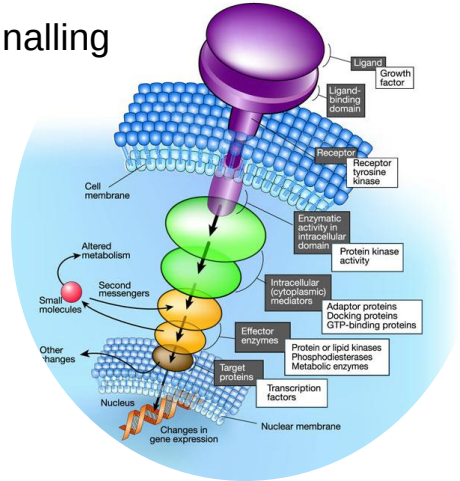
Genome, genes, proteins



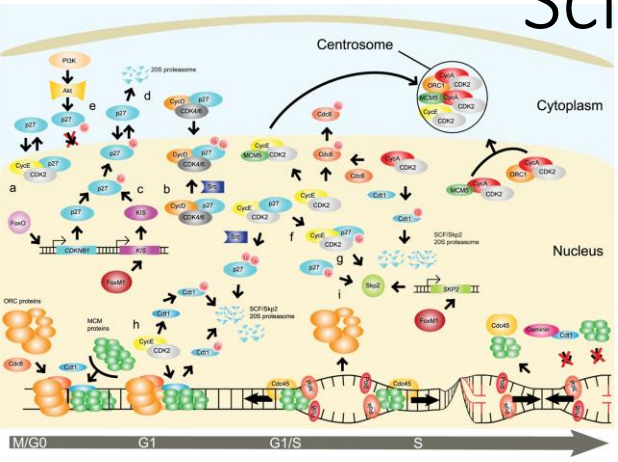
Gene expression regulation

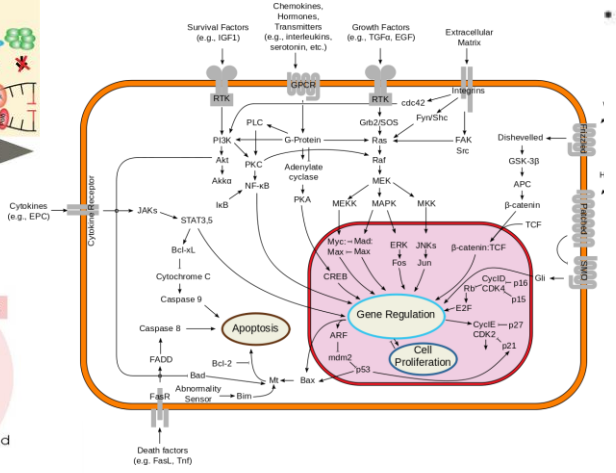


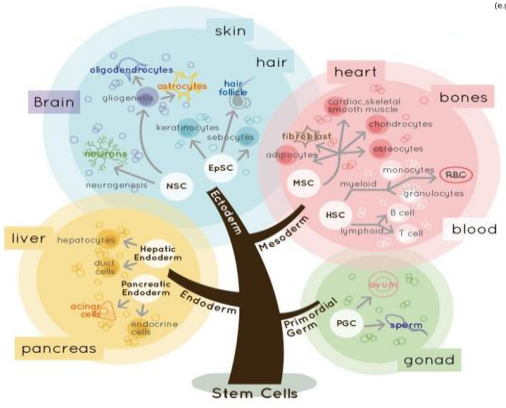
Cell Signalling



Scheme for non biologist

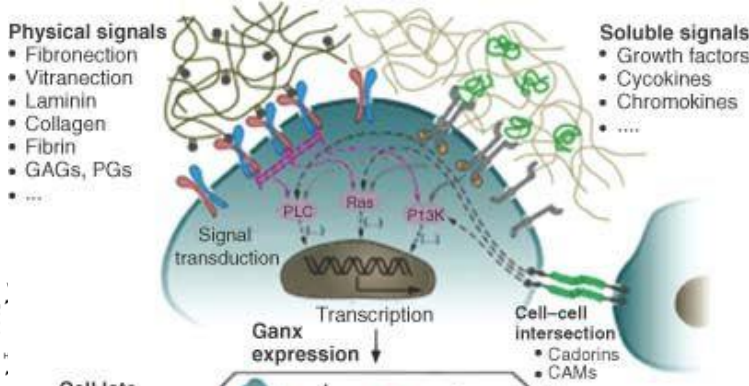


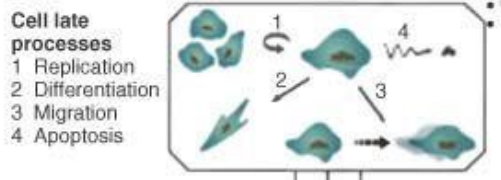


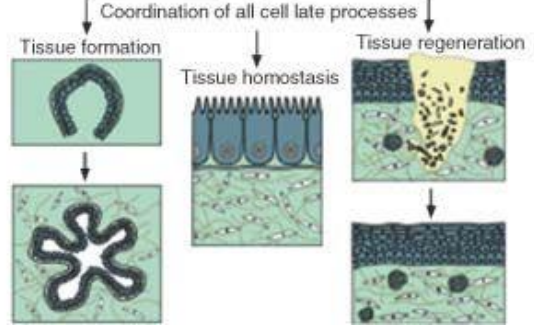


The tree of stem cell differentiation

Renew Differentiate Specialize







System biology depends on the type of cell, its processes and the surroundings

What are BioDBs like?

Characteristics of BioDBs

- High complexity
- Large amount and variability of information
- Multiple sources of information
- Custom or standardized format
- Multiple interpretations
- Unpredictable queries

Aspects to take into account

Resources providers

- Centers and organizations should commit to mantein the DBs

BioDBs

- There is a lot of variety and contains diverse information

Tools

- To find information in the DB
- To contrast sequences against BD
- To export the information

How can I look for information into the DBs

The information can be accessed through

- **key words, accession numbers, authors, etc...**

Homology searches

- **Sequences (strings) that are similar to others can be assessed**

Patterns searches

- **Sequences have patterns that can be recognised and stored serves to perform classifications, etc**

Prediction

- **Sequences that are alike to other sequences can be classified and predict biological functions, behaviour, species patterns specificity, etc**

Aspects to take into account

Resources providers

- Centers and organizations should commit to mantein the DBs

BioDBs

- There is a lot of variety and contains diverse information

Tools

- To find information in the DB
- To contrast sequences against BD
- To export the information

Which are the desired features of a BioDB?

- Organise the information efficiently
- Information must be accurate, verified, updated, cured
- Recover information easily, fast and accurately (efficient search engines)
- Update and revise the contents periodically
- Allow the incoming of new information (secuenses, annotations, etc)
- Share or transfer information to other DBs (related DBs)
- Share or transfer information to programs of analysis
- The format of the data should be standard*, universal to be compatible with other DBs and programs
- The content must be public and free
- Custom and friendly interphase for the user
- Accept user questions

The format of the information

How data is organized?

Database structure

- **Common elements in DBs structure:**
  - **Record**: auto consistent dataset. Ex: sequence of a gene and its features (metadata)
  - **Fields**: They are the elements that must fill each record
  - **Indexes**: They are lists ordered by field (precalculated).
  - **Relations**: How the information is connected with other in the same or other DBs.
  - **Queries**: Input accepted to make questions

Type of Databases

BioDBs can be classified according to many aspects:

- The aim of the DB → Depend on the purpose and the questions that are going to be made
- Confidence in the data → Depend on the level of curation of the DB
- Redundancy of the data → Depend on repetitive is the information in the DB
- Processing degree of the data → After processing and data assessment different information can be obtained and used as the main information of the DB

Type of Databases

According to the confidence in the data

Curated database (storage).

- Cured: The data is biologically confirmed. Even so, this does not mean that they can vary.

Non-curated databases

- The data was not confirmed. They are obtained from automatic processes or are partial results.

Type of Databases

According to the redundancy of the data

Redundant database (storage).

- It cannot be confirmed that the existing data is not repeated multiple times.

Non-redundant databases

- The data is not repeated under a certain criterion

Type of Databases

According to the processing degree of the data

Primary database (storage).

- They contain "raw" information obtained from experiments (usually laboratory)

Secondary databases

- Result derived from the bioinformatic analysis of the data stored in the primary DBs

La pregunta clave en bases de datos biologicas es:

¿Qué queremos almacenar?

Datos Biológicos

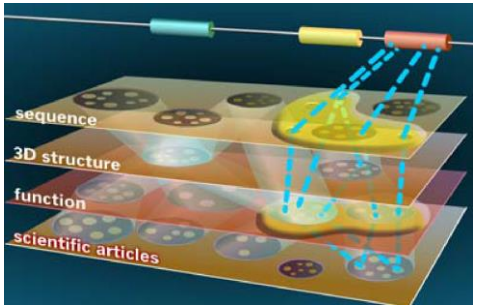
The main data can be of five types:

1. Organisms and species taxonomy
2. biological sequences
3. structural data
4. functional data
5. bibliography

Databases and information among BioDBs must be related because everything in biology is related

Each category of data has its own structure and requirements, which have a decisive influence on the design of databases.

The various types of data are **closely related to each other**: DNA coding sequences give rise to proteins with a three-dimensional structure and characteristic function; very often, proteins do not work alone but are part of metabolic pathways in which they establish important relationships with other types of biomolecules and, furthermore, all this information is conveniently reflected in scientific publications.



Main Biological Databases

Many of these databases are hosted on the websites of government or private centers that have created a uniform graphical environment that brings together a large number of databases.

- The National Center for Biotechnology Information, **NCBI** (http://www.ncbi.nlm.nih.gov/)
- The European Bioinformatics Institute, **EBI** (http://www.ebi.ac.uk)
- DNA Data Bank of Japan, **DDBJ** (https://www.ddbj.nig.ac.jp/index-e.html)
- The Switzerland Institute of Bioinformatics, **SIB** (http:/www.isb-sib.ch/)

Public biological Databases

Main Biological Databases

International Nucleotide Sequence Database Collaboration (INSDC)

Values

- open access for all
- globally comprehensive
- spanning life science domains
- permanent database of record
- public forum for the scientific process

Organisation

- established early 1980s
- major ongoing investment
- structure and governance
- model for scientific collaboration

http://www.insdc.org/

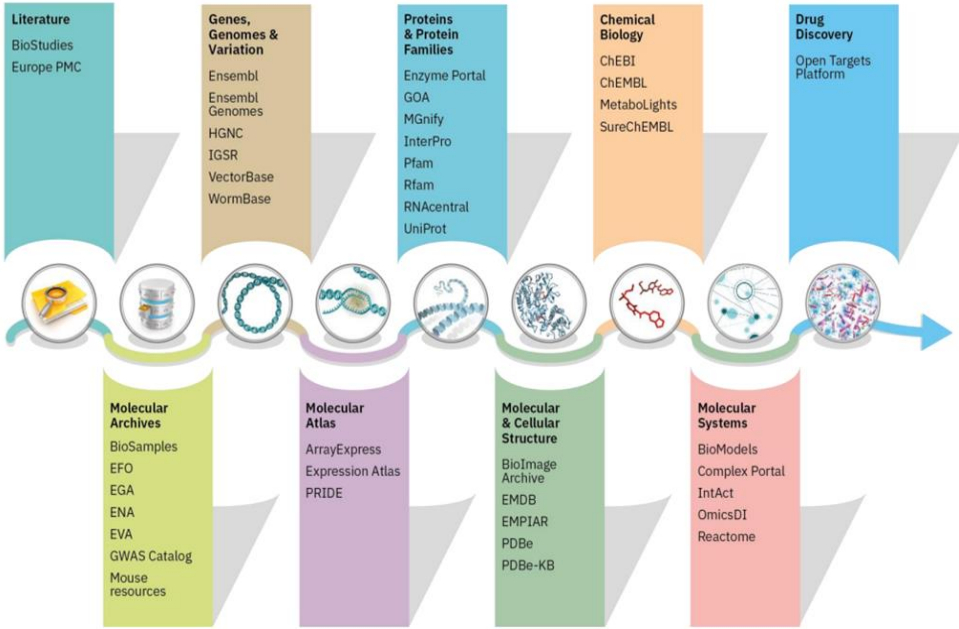
Instruments

- regular data exchange
- accession scheme
- data standards
- mandatory submission agreement
- services and software (node-level)

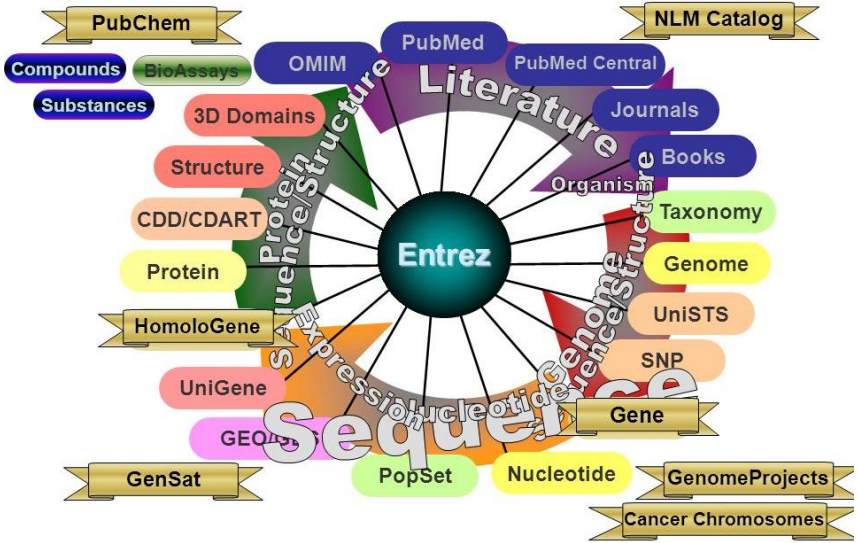
Public biologic data resources

Main Biological Databases

Data resources at EMBL-EBI

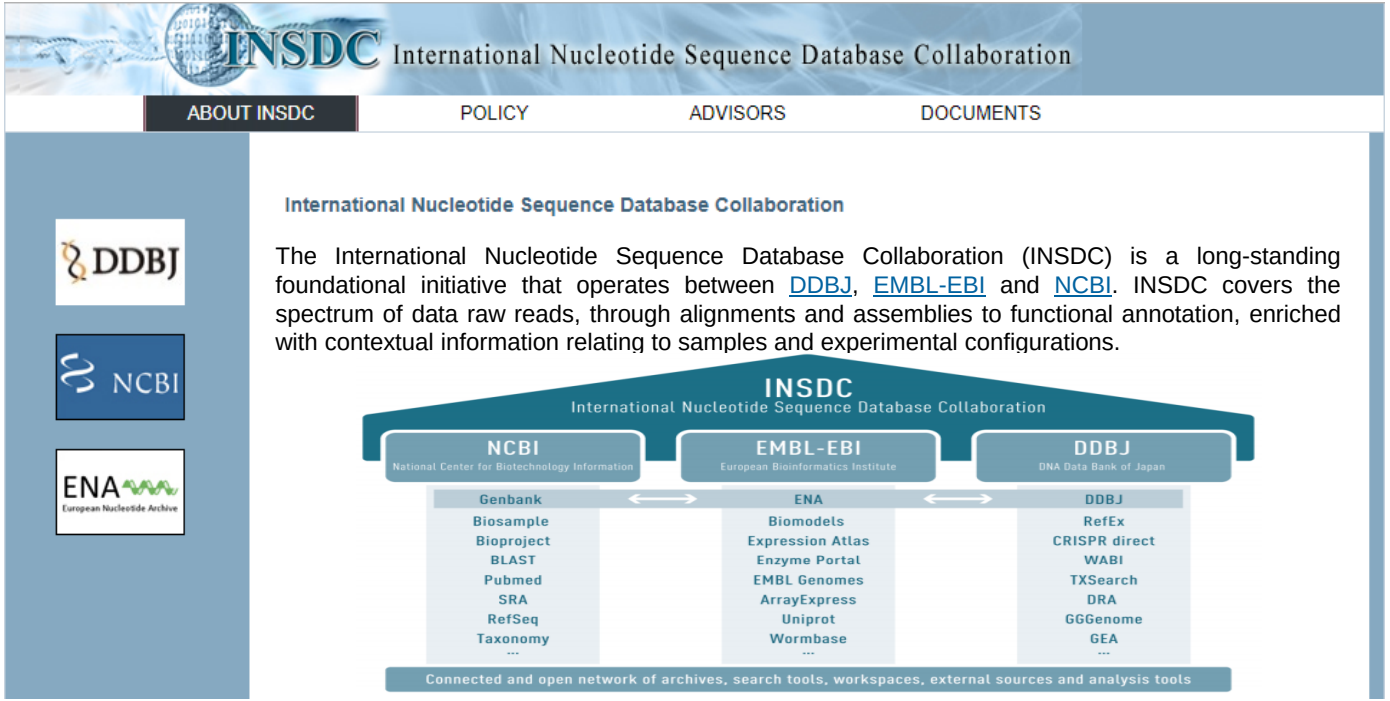


Data resources at NCBI-Entrez



Main Biological Databases

GenBank is managed and distributed by the NCBI (National Center for Biotechnology Information) in the United States. Together with the ENA (European Nucleotide Archive) and the DDBJ (DNA Data Bank of Japan), it forms the INSDC (International Nucleotide Sequence Database Collaboration) consortium. Every day, the three DBs update their contents so that they all have the same information.



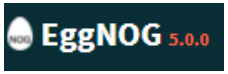
They establish connections between the different databases that allow all the information related to a specific biomolecule to be obtained easily and quickly. For example, the NCBI offers a platform that searches for information in 39 databases at once and allows easy "jumping" from one database to another (http://www.ncbi.nlm.nih.gov/gquery/) .

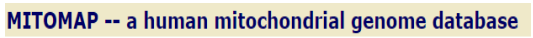
Public biological Databases

Main Biological Databases

1. Easily retrieving sequences from public databases
2. Downloading sequences and data/metadata from public databases
3. Construction of customized databases of nucleotides and amino acids: genes, CDSs, proteins, orthologs, genomes, transcriptomes, patterns or domains, genetic variants, etc
4. mySQL/SQL databases

- Fast access to databases
- Quick search and retrieving data and metadata
- Genes classification and prioritization
- Contaminant assessment
- Functional genomics
- comparative genomics
- etc





OXFORD ACADEMIC | Journals

You are here: NAR Journal Home » Database Summary Paper Categories

NAR Database Summary Paper Category List

[Nucleotide Sequence Databases](#)
[RNA sequence databases](#)
[Protein sequence databases](#)
[Structure Databases](#)
[Genomics Databases (non-vertebrate)](#)
[Metabolic and Signaling Pathways](#)
[Human and other Vertebrate Genomes](#)
[Human Genes and Diseases](#)
[Microarray Data and other Gene Expression Databases](#)
[Proteomics Resources](#)
[Other Molecular Biology Databases](#)
[Organelle databases](#)
[Plant databases](#)
[Immunological databases](#)
[Cell biology](#)

Nucleic Acids Research, 2020, Vol. 48, Database issue **D1–D8**
doi: 10.1093/nar/gkz1161

The 27th annual Nucleic Acids Research database issue and molecular biology database collection

Daniel J. Rigden[1,*] and Xosé M. Fernández[2]

[1]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [2]Institut Curie, 25 rue d'Ulm, 75005 Paris, France

Each initial year there is an editorial dedicated to new Biological Databases

Growth rate of BioDBs

The International Nucleotide Sequence Database Collaboration

Guy Cochrane[1,*], Ilene Karsch-Mizrachi[2], Toshihisa Takagi[3] and International Nucleotide Sequence Database Collaboration

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [3]DDBJ Center, National Institute for Genetics, Mishima, Japan

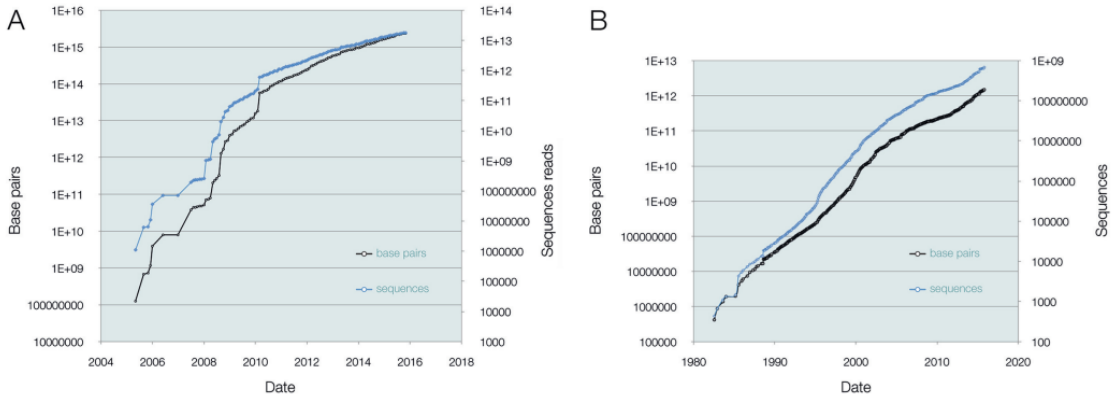
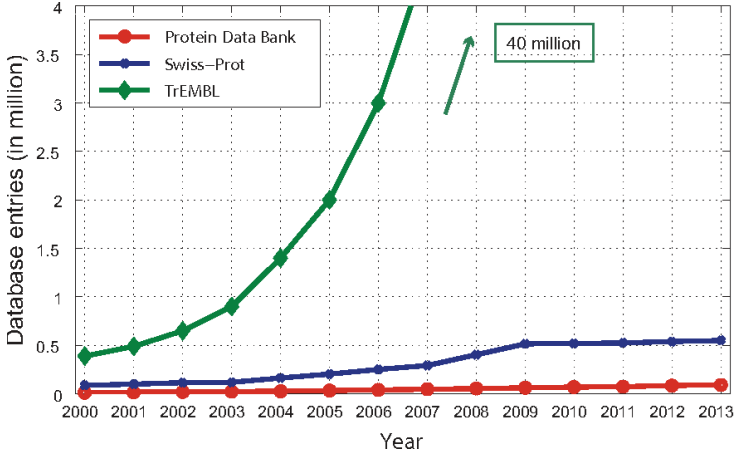


Figure 1. Cumulative growth in INSDC. (**A**) Base pairs (black, 2365.5 trillion) and sequence reads (blue, 17.8 trillion) for INSDC raw data. (**B**) Base pairs (black 1449 billion) and sequences (blue, 651.5 million) in INSDC assembled/annotated data.



Growth rate of BioDBs of Protein structures

Protein Data Bank: the single global archive for 3D macromolecular structure data

wwPDB consortium

Received September 14, 2018; Revised September 28, 2018; Editorial Decision October 01, 2018; Accepted October 05, 2018

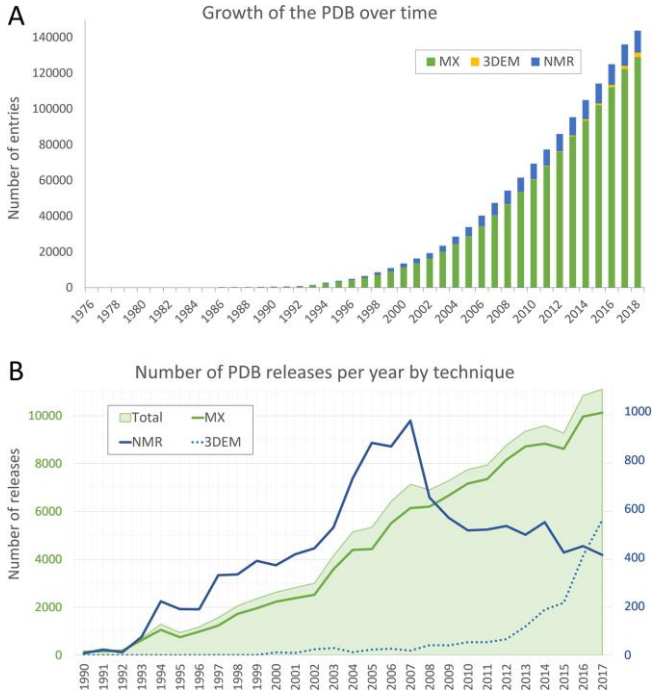


Figure 1. (**A**) Growth the PDB Core Archive. Total height of each bar indicates aggregate released structures, coloured by experimental technique (MX—green, 3DEM—yellow, NMR—blue). (**B**) Number of PDB structures released annually. All PDB Core Archive structures are indicated with light green shading, and MX structures are shown with a solid green line, plotted with respect to the green primary axis (left). NMR structures (blue solid line) and 3DEM structures (blue dashed line) are plotted with respect to the blue secondary axis (right).

Nucleotide and protein databases

EMBL-EBI

NCBI

Gene-Bank (NCBI nucleotide) http://www.ncbi.nlm.nih.gov/genebank

- Nucleotide DB belonging to NCBI (NIH)
- Includes: mRNA, genomic DNA (1 or multiple genes), coding regions, ribosomal
- RNA Primary information file. (NOT cured)
- The person responsible is the author
- Multiple entries for each loci. (SNP detection)
- If it corresponds to CDS, an ID is assigned that refers to the Protein data base (NBCI) and Uniprot (TrEMBL)
- Participate in the INSDC

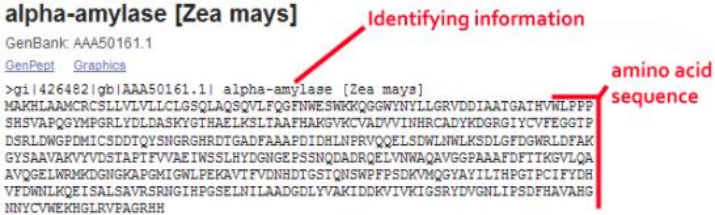
Gene-Bank (NCBI nucleotide) http://www.ncbi.nlm.nih.gov/genebank

- Redundant (it is a Bank, it does not seek to unify data) and has errors
- Difficult to update in order to correct, improve and keep updated the annotation of the records,
- NCBI created RefSeq (curated collection of GenBank records) takes records from GenBank and updates/corrects them unifies to reduce redundancy
- Accession numbers of type XX_123456
- Today for NGS data, there is SRA

- **Record: file in text format (plain-text)**
- **Subdivided into: Sections divided in Fields (Feature Table)**
- **The sequence itself is saved in Fasta Format:**

  >Definition Line

  Sequence 1 letter code (Max/Min) 60 characters per line.



And then the records have different sections:

Header (specific to each DB)

The Fields (defined in the Feature Table):

LOCUS(GB) ID(EMBL): Contains the unique identification of the record

Definition lines: contain summarized Biological information.

Taxonomy: Taxonomic Information

Reference: At least 1 reference

Comments:

CrossRef: Cross-reference fields to other databases.

Overview of the Feature Table format

Record of GB/EMBL/INSDC

The overall goal of the feature table design is to provide an extensive vocabulary for describing features in a flexible framework for manipulating them.

The Feature Table documentation represents the shared rules that allow the three databases to exchange data on a daily basis.

The range of features to be represented is diverse, including regions which:

- perform a biological function,
- affect or are the result of the expression of a biological function, interact with other molecules,
- affect replication of a sequence,
- affect or are the result of recombination of different sequences, are a recognizable repeated unit,
- have secondary or tertiary structure,
- exhibit variation, or have been revised or corrected.
- etc

nBank and EMBL divisions

GenBank and EMBL divisions:

sequences are distributed in 20 divisions:

•12 of them are taxonomic

•8 are functional, as they refer to the various sequencing strategies

Taxonomic Divisions		Functional Divisions	
Division	**Description**	**Division**	**Description**
SYN	Synthetic	**TSA**	Transcriptome shotgun data
PHG	Phages	**WGS**	Whole-genome shotgun data
ENV	Environmental samples	**PAT**	Patented sequences
VRL	Viruses	**GSS**	Genome survey sequences
BCT	Bacteria	**EST**	Expressed sequence tags
PLN	Plants	**HTG**	High-throughput genomic
MAM	Other mammals	**STS**	Sequence tagged sites
VRT	Other vertebrates	**HTC**	High-throughput cDNA
PRI	Primates		
UNA	Unannotated		
ROD	Rodents		
INV	Invertebrates		

Sequence division		Database
Organismal		
BCT	Bacterial	DDBJ, GenBank
PRO	Prokaryotic	EMBL
FUN	Fungal	EMBL
HUM	Human	DDBJ, EMBL
PRI	Primate	DDBJ, EMBL, GenBank
ROD	Rodent	DDBJ, EMBL, GenBank
MAM	Other mammalian	DDBJ, EMBL, GenBank
VRT	Other vertebrate	DDBJ, EMBL, GenBank
INV	Invertebrate	DDBJ, EMBL, GenBank
PLN	Plant	DDBJ, EMBL, GenBank
ORG	Organelle	EMBL
VRL	Viral	DDBJ, EMBL, GenBank
PHG	Phage	DDBJ, EMBL, GenBank
RNA	Structural RNA	DDBJ, EMBL, GenBank
SYN	Synthetic and chimeric	DDBJ, EMBL, GenBank
UNA	Unannotated	DDBJ, EMBL, GenBank
Functional		
EST	Expressed sequence tag	DDBJ, EMBL, GenBank
STS	Sequence tagged site	DDBJ, EMBL, GenBank
GSS	Genome survey	DDBJ, EMBL, GenBank
HTG	High-throughput genomic	DDBJ, EMBL, GenBank
PAT	Patent	DDBJ, EMBL, GenBank
CON[a]	Virtual contigs of segmented sequences	DDBJ, EMBL, GenBank

[a]This division, which will appear in future database releases, is designed to contain instructions for assembly of segmented sequence records.

y Annotation

- The Third Party Annotation (TPA) database contains sequences already existing but annotated by new authors from published data.
- Two types of records: Experimental Annotations (evidence of wet results) and Inferential
- (The annotation is BioInfo analysis results and awaits confirmation exp.)
- TPA connects GenBank with RefSeq as it allows adding annotations supported by new evidence from public records.
- The records are distinguished by the fact that the fields begin with "TPA"
- The TPA/GB records ratio is 1/10000 approx.

Redundancy and biological significance
The RefSeq Database

GenBank sequence records are owned by the original submitter and cannot be altered by a third party. RefSeq sequences are not part of the INSDC but are derived from INSDC sequences to provide non-redundant curated data representing our current knowledge of known genes

RefSeq

Ensembl

GENCODE

- Genbank and EBI are redundant (it is a Bank, it does not seek to unify data)
- They have errors
- Difficult to update in order to correct, improve and keep updated the annotation of the records
- NCBI created RefSeq (curated collection of GenBank records) takes records from GenBank and updates/corrects them unifies to reduce redundancy
- EBI-EMBL feed Ensembl and gencode and cure specific datasets

RefSeq: DB of non-redundant and highly annotated and curated DNA, RNA and Protein sequences.

- It was born in 2003 with the aim of avoiding the problems associated with the redundancy of GB with the aim of representing the sequence considered the "standard" or "normal", with a large number of annotations.
- Definition of NO-redundancy consists of providing "one" example of each molecule per organism
- Limited to a set of model organisms for which there is sufficient data in the primary databases.
- The records are similar to those of GB but the UID has a format of 2(letters) + 6(numbers), according to:

NT: genomic DNA from Contigs

NM: mRNA, sequenced/verified experimentally

NP: Experimentally sequenced/verified proteins

XN: in-silico predicted mRNAs

XP: in-silico Predicted Proteins

RefSeq: BD of non-redundant and highly annotated and curated DNA, RNA and Protein sequences.

NCBI created RefSeq (curated collection of GenBank records) takes records from GenBank and updates/corrects them and unifies to reduce redundancy

RefSeq is the NCBI database of reference sequences; a curated, non-redundant set including genomic DNA contigs, mRNAs and proteins for known genes, and entire chromosomes.

Similar format GeneBank

- **Access via NCBI**
- **You can think of the following analogy:**
  - GeneBank is the paper
  - RefSeq is the review

GenBank	vs	RefSeq
Non-cured		Cured

How Genbank and EMBL works

How Genbank and EMBL works

The sequences are sent directly electronically

- Each GenBank record contains an uninterrupted sequence of a polynucleotide molecule. We can find several types of polynucleotides: genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA or small cytoplasmic RNA.
- The minimum size of the sequences stored in GenBank is 50 nucleotides
- The records include bibliographic and biological annotations.
- GenBank staff assign an accession number to the record containing the sequence and annotations.
- The access number is a unique identifier used by the three databases (GenBank, ENA and DDBJ) and will always be associated with that sequence (a combination of letters and numbers, such as U12345).
- If changes are made to the sequence or log entries, what does change is the version of the sequence, which is indicated after the accession number (for example: U12345.1).
- The NCBI assigns each version a unique identifier called "gi" (GenInfo Identifier).
- All of this information appears on the two lines of the log beginning with the words ACCESSION and VERSION. Example:
- ACCESSION U12345
- VERSION U12345.1 GI: 7654321

Structure of a GenBank Record

- **Header:** This is the part of the record where the database staff is most involved and where it is possible to find slight variations between GenBank and the other databases of the INSDC consortium. Contains general information about the registry, spread over several lines of information.
- **LOCUS:** Name of the genetic locus where the sequence resides, sequence length, type of molecule, division of GenBank and date of the last modification.
- **DEFINITION:** Organism where it comes from, name of the gene or protein, brief description of its function. This is the same line that appears in FASTA format after the ">" symbol.
- **ACCESSION:** Access number. It is associated with the registry forever, even if it is modified. It is the one cited in the publications.
- **VERSION:** The version number changes each time a modification is made. Every version has the same accession number, but a GI (GeneInfo Identifier) different to be able to have a history of the changes that the sequence.
- **KEYWORDS** Keywords.
- **SOURCE:** Common name and scientific name of the organism from which the sequence comes.
- **ORGANISM:** Complete taxonomy of the organism from which the sequence comes.
- **Bibliographic references:** Each record contains at least one bibliographic reference that includes the name of the authors, the title of the article, the journal where it has been published and the PubMed identifier (PMID). When there is more than one, they are numbered and displayed in chronological order, starting with the oldest. The last reference contains information about the authors who have submitted the sequence to GenBank and the date of submission.
- **REFERENCE:** They appear numbered and in chronological order, starting with the oldest. The name of the authors, the title of the article, the journal that published it and the PUBMED identifier (PMID) are included. The last reference contains information about the authors who have submitted the sequence to GenBank.
- **COMMENTS:** This line is optional. If the record has been modified, links to previous versions can be included here.

NCBI

Structure of a EMBL-EBI Record

The flatfile format used by the EMBL to represent database records for nucleotide and peptide sequences from EMBL database (Stoesser et al., 2002). The EMBL flat file comprises of a series of strictly controlled line types presented in a tabular manner and consisting of four major blocks of data:

- Descriptions and identifiers.
- Citations: citation details of the associated publications and the name and contact details of the original submitter.
- Features: detailed source information, biological features comprised of feature locations, feature qualifiers, etc.
- Sequence: total sequence length, base composition (SQ) and sequence.

EMBL-EBI

enBank and EMBL records

NCBI

EMBL-EBI

The Flatfile Format

```
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
  JOURNAL   DNA Seq. 14 (4), 265-277 (2003)
REFERENCE   2  (residues 1 to 348)
  AUTHORS   Waldbieser,G.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-OCT-1997) Catfish Genetics Research Unit, USDA -
            Agricultural Research Service, 141 Experiment Station Road,
            Stoneville, MS 38776, USA
REFERENCE   3  (residues 1 to 348)
  AUTHORS   Waldbieser,G.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-FEB-2002) Catfish Genetics Research Unit, USDA -
            Agricultural Research Service, 141 Experiment Station Road,
            Stoneville, MS 38776, USA
COMMENT     Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source          1..348
                     /organism="Ictalurus punctatus"
                     /organelle="mitochondrion"
                     /strain="Norris"
                     /db_xref="taxon:7998"
     Protein         1..348
                     /product="NADH dehydrogenase subunit 2"
     CDS             1..348
                     /gene="ND2"
                     /coded_by="AF482987.1:4903..5949"
                     /transl_table=2
ORIGIN
        1 mspyvitill sslglgtalt fmsshwllaw agleintlai lplmaqhhhp ravesttkyf
       61 laqaaaaati lfastinawt tgewniycls hpsatilita alalkvglap vhfwappvwq
      121 gltlttglim atwgklapfa liiqmapfth pilltlgll svfiggwggl nqtqlkkila
      181 yssiahlgvm iivtqykpql tvlvlityii mtsatfltfk lmattktintl amswakvpti
      241 tamaalalls lgglppltgf mpkwliiqel tmqglpltat mmtlsallsl yfylrlcyam
      301 tltispntnn ssapwrlqnt qataplatlm imtlllplt plaqtltn
//
```

Header

Feature Table

Sequence

```
ID   RNGTPCHI   standard; RNA; ROD; 1016 BP.
XX
DT   01-AUG-1991 (Rel. 28, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 2)
XX
DE   Rat GTP cyclohydrolase I mRNA, complete cds.
XX
KW   GTP cyclohydrolase I.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
XX
RN   [1]
RP   1-1016
RX   MEDLINE; 91093270.
RX   PUBMED; 1985963.
RA   Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;
RT   "Cloning and sequencing of cDNA encoding rat GTP cyclohydrolase I: The
RT   first enzyme of the tetrahydrobiopterin biosynthetic pathway";
RL   J. Biol. Chem. 266(2):765-769(1991).
XX
FT   CDS             128..853
FT                   /codon_start=1
FT                   /db_xref="GOA:P22288"
FT                   /db_xref="SWISS-PROT:P22288"
FT                   /EC_number="3.5.4.16"
FT                   /gene="GTP cyclohydrolase I"
FT                   /product="GTP cyclohydrolase I"
FT                   /protein_id="AAA41299.1"
FT                   /translation="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAUK
FT                   AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI
FT                   SDVLNDAIFDEDHDEMVIVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV
FT                   EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRGVQKMNSKTVTSTML
FT                   GVFREDPKTREEFLTLIRS"
SQ   Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
     gacttcgaac ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggccg        60
     ccggctaagc cgagccgcag cgcttggtag caccttaggg tgtctcggga gcaatcgcgc       120
     cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct       180
        ...
     catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg       900
     ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttatttctca       960
     ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caactt          1016
//
```

Molecule type

Name

Date of creation and last update

Free text description

Keywords describing the molecule

Organism

Article the sequence was published in

Structural annotation (coding sequence)

Link to functional annotation of resulting protein

Translated protein sequence

Sequence of bases

How to search for sequences in the DB?

NCBI-ENTREZ – EBI-EMBL

- HIGHLY “INDEXED” GOOGLE-LIKE SEARCH ENGINE
- ACCESS TO INSDC, PROTEINS, PUBMED ETC. WITH THE SAME BROWSER
- LINKS-OUT TO OTHER DATABASES (UNIPROT, PDB)
- SEQUENCE SEARCH BY ALIGNMENT
- SEARCH BY KEYWORD
- OTHERS

ENTREZ

- IT IS NOT A DATABASE.
- IT IS AN INTERFACE THROUGH WHICH ALL THE NCBI DATABASES CAN BE ACCESSED/BROWSED IN AN INTEGRATED MANNER.
- INTEGRATION IS ACHIEVED THROUGH THE CONCEPTS OF NEIGHBORHOODING (NB) AND HARD-LINKS (HL).
- HL: ARE CONNECTIONS THAT HAVE RECORDS FROM ONE DATABASE TO ANOTHER.
- EXAMPLE THE LINK TO GENEPEPT FROM GENEBANK.
- NB: THEY ARE RELATIONSHIPS BETWEEN THE RECORDS OF THE SAME DATABASE.
- IT IS DONE BY COMPARING THE SEQUENCE USING BLAST
- COMPARING STRUCTURE USING VAST
- COMPARING TEXT FIELDS USING WEIGHTED KEY TERMS
- IT IS AN ALGORITHM THAT COMPARES TEXTS BASED ON THE WORDS THEY SHARE AND THEIR RELATIVE ORDER OR CONTEXT.

ENTREZ

TAXONOMY

Was this helpful?

Homo sapiens

Homo sapiens (human) is a species of primate in the family Hominidae (great apes).

Taxonomy ID: 9606

Genomes
Browse and download

Genes
Browse and download

Genome Data Viewer
Browse the reference genome

BLAST
Search the reference sequence

Literature	
Bookshelf	168,486
MeSH	17,381
NLM Catalog	71,014
PubMed	21,039,850
PubMed Central	5,236,534

Genes	
Gene	6,073,060
GEO DataSets	2,723,499
GEO Profiles	64,608,705
HomoloGene	18,907
PopSet	41,982

Proteins	
Conserved Domains	3,506
Identical Protein Groups	2,956,008
Protein	80,496,576
Protein Family Models	15,872
Structure	71,790

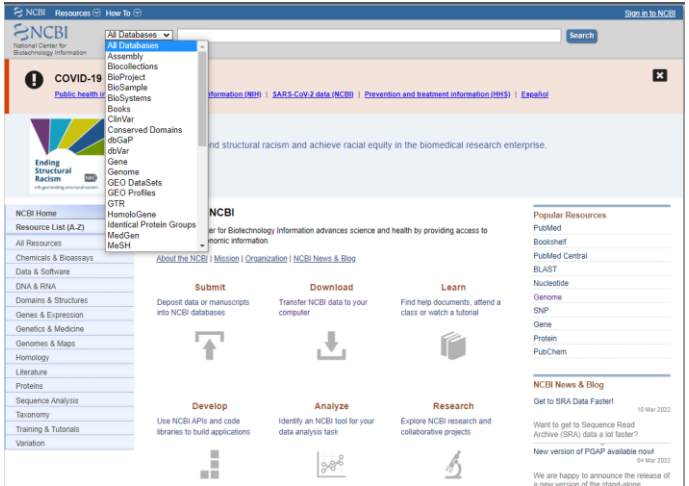
Genomes	
Assembly	2,110
BioCollections	7
BioProject	106,914
BioSample	11,284,353
Genome	1,735
Nucleotide	52,368,913
SRA	6,229,806
Taxonomy	1

Clinical	
ClinicalTrials.gov	47,498
ClinVar	1,199,829
dbGaP	814
dbSNP	1,071,975,857
dbVar	7,400,967
GTR	19,728
MedGen	787
OMIM	18,082

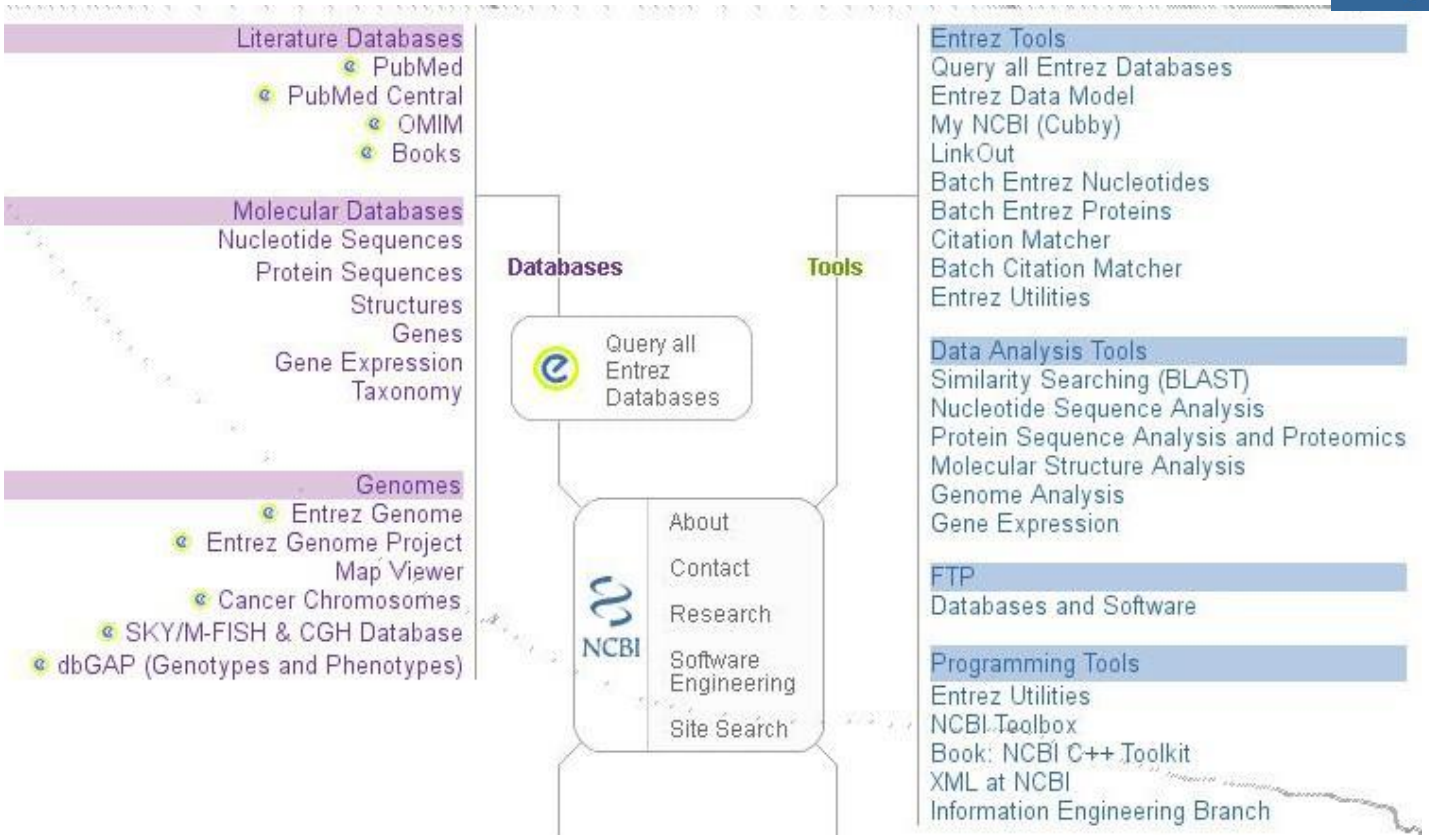
PubChem	
BioAssays	639,426
Compounds	225,125
Pathways	52,762
Substances	225,119

Sistemas de búsqueda avanzada: Gquery

- It is a user interface.
- It constitutes the link between the user and the databases.
- It allows making simple queries and obtaining results, even without knowing the architecture of the databases.
- However, knowing that architecture can make searches more accurate and efficient.
- It is not easy to access this knowledge and its use is not very intuitive, so it is always recommended to visit the NCBI help.



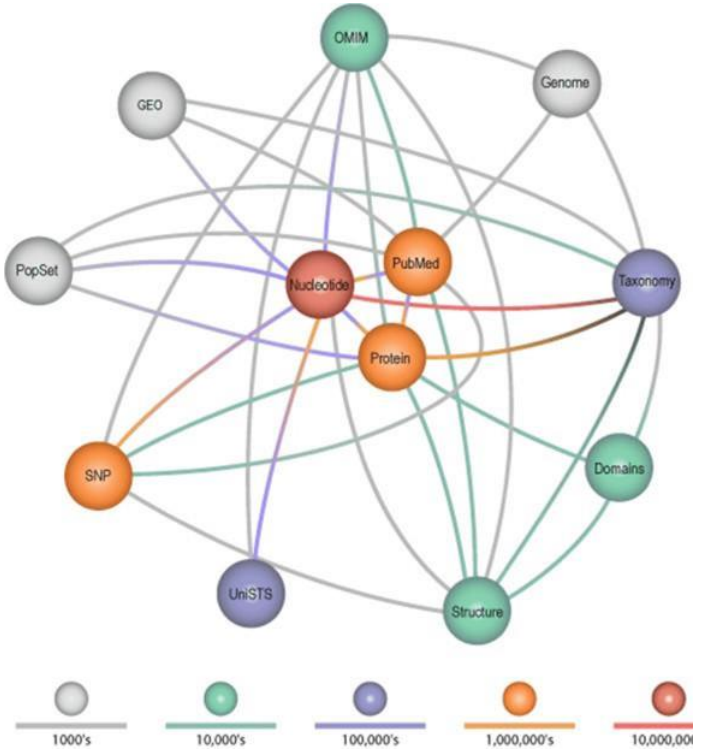
NCBI structure



NCBI structure

Functioning of the Entrez system is based on the connections between nodes that correspond to specific databases. The original system, which contained three nodes, i.e. GenBank (Nucleotide), Proteins and PubMed, has evolved in recent years, adding new nodes, including [4]:

- **Taxonomy [8],** organized around the names and phylogenetic relationships between organisms;
- **Structure [9],** organized around the three-dimensional structures of proteins and nucleic acids;
- **Genome [10],** representing completely sequenced organisms and those for which sequencing is in progress, together with links to genomic data available for these organisms;
- **PopSet [5],** a set of DNA sequences that have been collected to analyze the evolutionary relatedness of a population;
- **OMIM [11],** which is a database of all known diseases with a genetic basis;
- **SNP (dbSNP) [12],** the Single Nucleotide Polymorphism database – a public-domain archive for a broad collection of simple genetic polymorphisms.



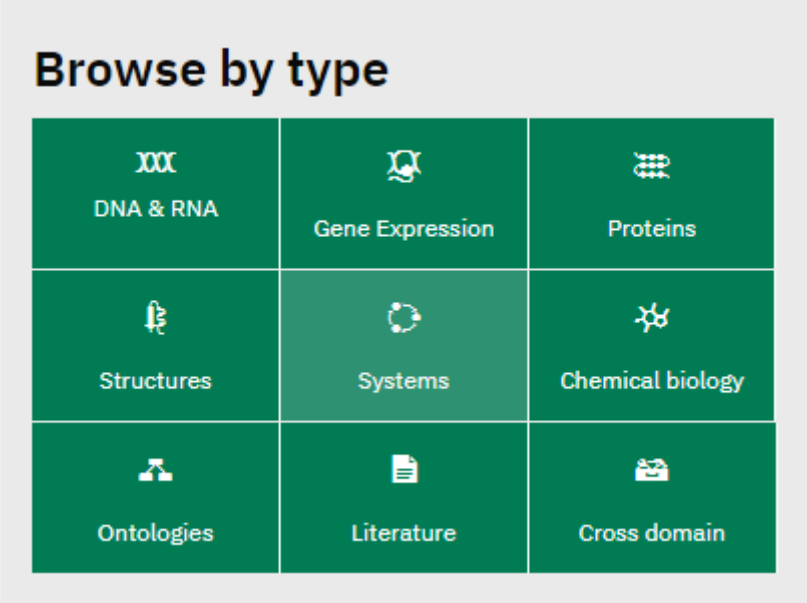
EBI search

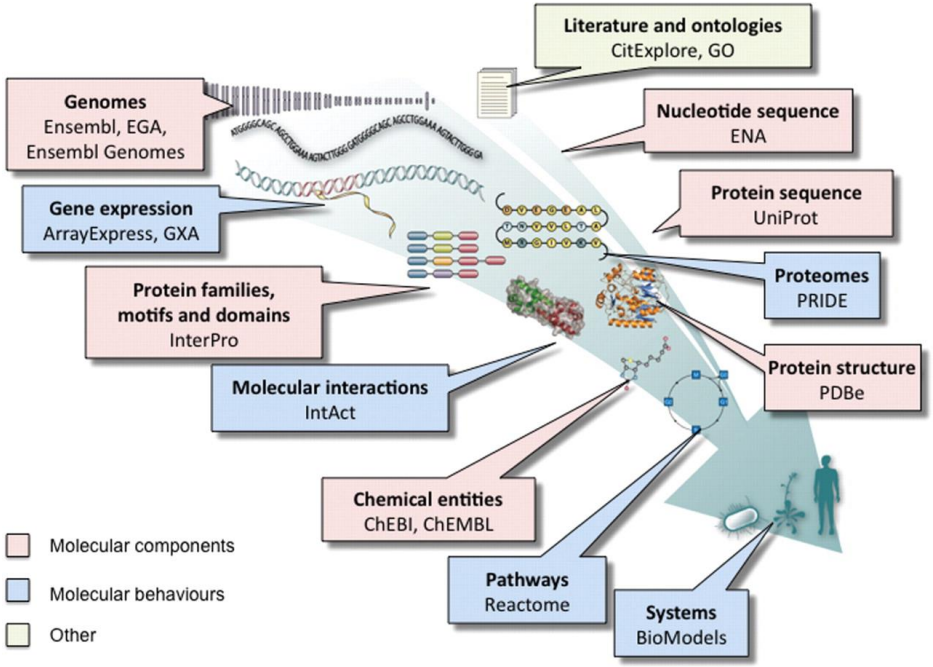
EMBL-EBI

- Ebi Search Is A Scalable Text Search Engine That Provides Easy And Uniform Access To The Biological Data Resources Hosted At The European Bioinformatics Institute (Embl-ebi).
- The Data Resources In Ebi Search Include: Nucleotide And Protein Sequences At Both The Genomic And Proteomic Levels; Structures Ranging From Chemicals To Macro-molecular Complexes; Gene-expression Experiments; Binary Level Molecular Interactions As Well As Reaction Maps And Pathway Models; Functional Classifications; Biological Ontologies; Diseases; And Comprehensive Literature Libraries Covering The Biomedical Sciences And Related Intellectual Property.
- Ebi Search, Based On Apache Lucene, Presents Search Results That Are Up-to-date With The Data Resources And Provides An Easy Inter-domain Navigation Via A Network Of Cross-references.

EMBL-EBI structure

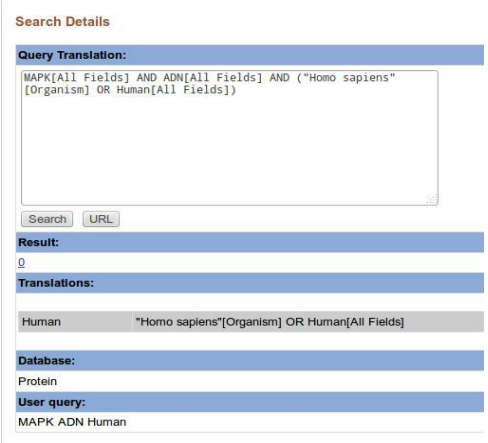
The European Nucleotide Archive and the protein sequence resource UniProt (then known as Swiss-Prot–TrEMBL) were the original EMBL-EBI databases. Since then, EMBL-EBI has played a major part in the bioinformatics revolution.

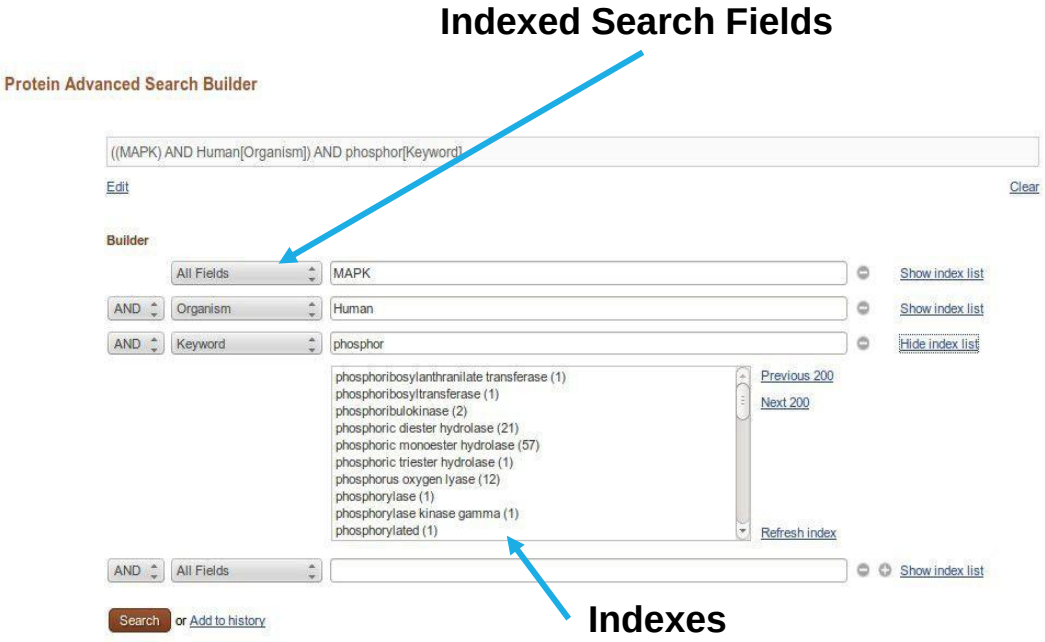




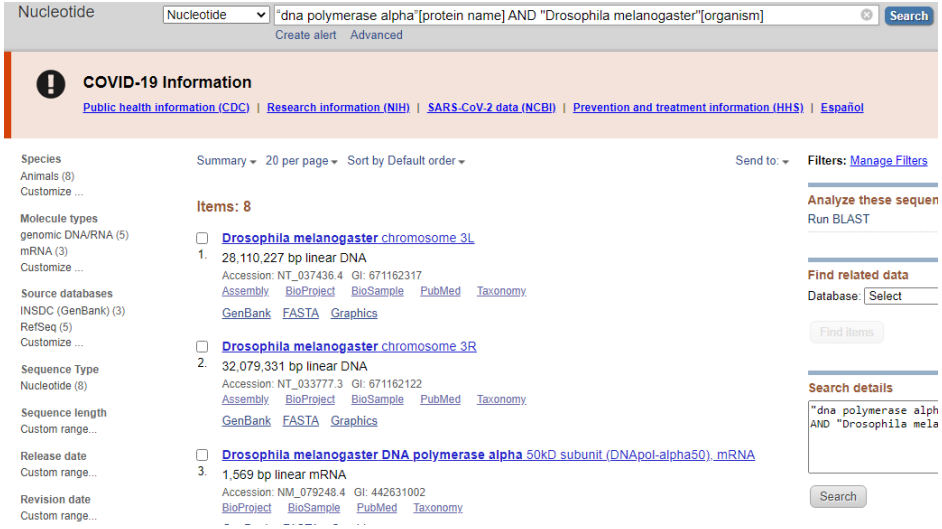
How to search sequences in the NCBI and EMBL-EBI Databases

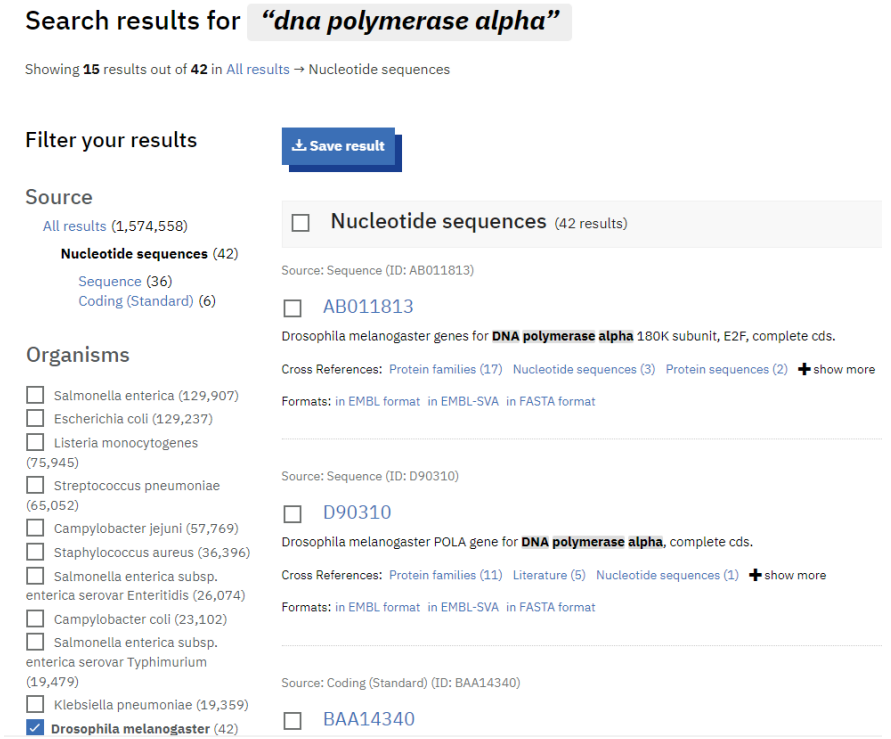
- Select the right database
- Know the fields/indexes and how to use them
- Use logical connectors (AND, OR and NOT) to perform combined searches
- Limit the search with filters
- Examples of things to consider...





How to search sequences in the NCBI and EMBL-EBI Databases





To download big datasets It's better to use tools (command line) to download the required data

Genome databases

Many specialized DBs

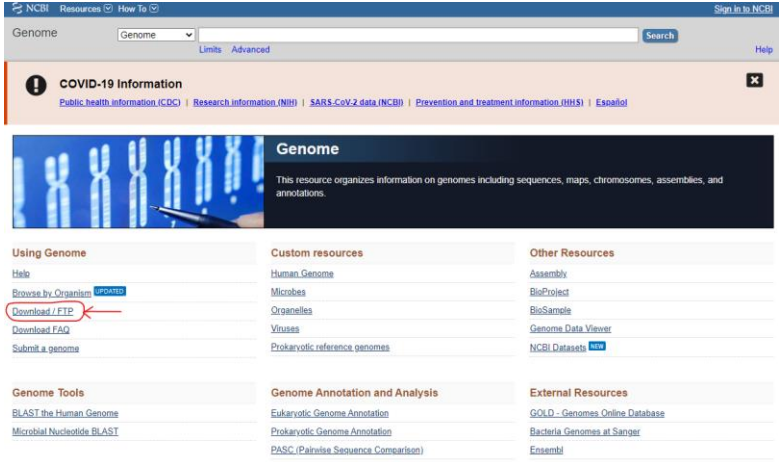
They are responsible for maintaining and updating the sequences and annotations of complete genomes.

- Ensembl (EBI-EMBL)
- Genome viewer (NCBI)
- Goldenpath (UCSC)

There are also specialized genomic resources

- Transfact: transcription factor binding sites
- EST: Expressed Sequence Tags
- UTRDB: Untranslated regions
- SpliceSitesDB: Splicing signal pairs

Downloading Genomes files from NCBI



Index of /genomes/refseq/vertebrate_mammalian/Homo_sapiens

Name	Last modified	Size
Parent Directory		-
all_assembly_versions/	2022-03-10 10:09	-
annotation_releases/	2022-03-12 12:44	-
latest_assembly_versions/	2022-03-10 10:09	-
reference/	2022-03-10 10:09	-
README.txt	2020-09-02 16:26	43K
annotation_hashes.txt	2022-03-10 00:35	3.5K
assembly_summary.txt	2022-03-10 01:12	706
assembly_summary_historical.txt	2022-03-10 01:45	7.5K

Index of /genomes

https://ftp.ncbi.nlm.nih.gov/genomes/

Name	Last modified	Size
Parent Directory		-
ASSEMBLY_REPORTS/	2022-03-12 05:56	-
CLUSTERS/	2017-12-04 10:38	-
GENOME_REPORTS/	2020-04-08 12:51	-
HUMAN_MICROBIOM/	2012-04-19 03:27	-
INFLUENZA/	2020-10-14 04:02	-
MapView/	2022-02-07 22:48	-
TARGET/	2017-10-23 11:48	-
TOOLS/	2020-05-26 10:19	-
Viruses/	2022-03-14 05:34	-
all/	2019-04-02 09:50	-
archive/	2020-06-12 15:55	-
genbank/	2022-03-11 20:30	-
refseq/	2022-03-12 05:52	-
README.txt	2020-01-27 16:55	11K
README_GFF3.txt	2020-01-06 13:00	35K
README_assembly_summary.txt	2021-10-28 13:05	15K
README_change_notice.txt	2016-09-22 15:57	6.6K

Index of /genomes/refseq/vertebrate_mammalian/Cricetulus_griseus

Name	Last modified	Size
Parent Directory		-
all_assembly_versions/	2022-03-10 11:21	-
annotation_releases/	2022-03-10 20:37	-
latest_assembly_versions/	2022-03-10 11:21	-
representative/	2022-03-10 03:18	-
README.txt	2020-09-02 16:26	43K
annotation_hashes.txt	2022-03-10 00:35	1.1K
assembly_summary.txt	2022-03-10 01:12	1.0K
assembly_summary_historical.txt	2022-03-10 01:45	1.0K

It's better to use tools (command line) to download the required data

Downloading Genomes files from NCBI

Let's suppose we want to download the proteins, genes or genome from human from a former assembly

GenBank ▾ Send to: ▾

⚠ Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly

NCBI Reference Sequence: NC_000001.11

FASTA Graphics

Go to: ⌄

```
LOCUS       NC_000001          248956422 bp    DNA     linear   CON 22-NOV-2021
DEFINITION  Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly.
ACCESSION   NC_000001
VERSION     NC_000001.11
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.39
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

GCF_000001405.39

Index of /genomes/all

Name	Last modified	Size
Parent Directory		-
GCA/	2022-03-07 01:23	-
GCF/	2022-03-04 05:11	-
annotation_releases/	2022-03-12 13:11	-
README.txt	2020-09-02 16:26	43K
README_assembly_structure.txt	2016-08-09 20:46	14K
README_change_notice.txt	2016-09-22 15:57	6.6K

Index of /genomes/all/GCF

Name	Last modified	Size
Parent Directory		-
000/	2017-08-29 23:21	-
001/	2018-11-16 07:42	-
002/	2020-07-28 11:58	-
003/	2022-02-15 16:04	-
004/	2021-08-19 01:18	-
005/	2021-08-19 01:19	-

Index of /genomes/all/GCF/000

Name	Last modified	Size
Parent Directory		-
001/	2016-09-20 10:01	-
002/	2017-09-08 11:08	-
003/	2021-12-22 00:13	-
004/	2017-03-17 21:53	-
005/	2016-09-20 10:01	-

Index of /genomes/all/GCF/000/001

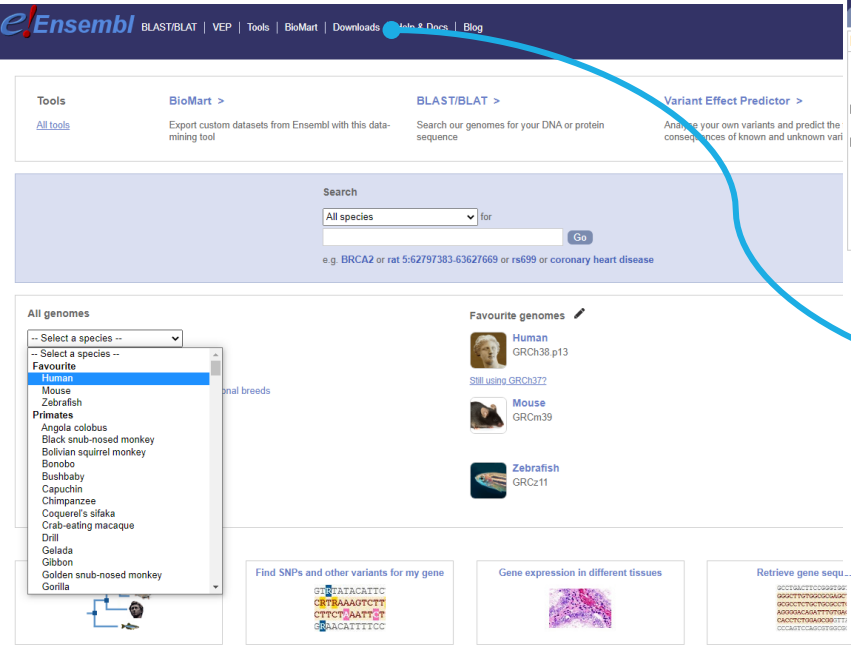
Name	Last modified	Size
Parent Directory		-
215/	2016-09-20 12:06	-
405/	2022-03-12 12:45	-
515/	2016-09-20 12:08	-
545/	2020-05-20 08:00	-

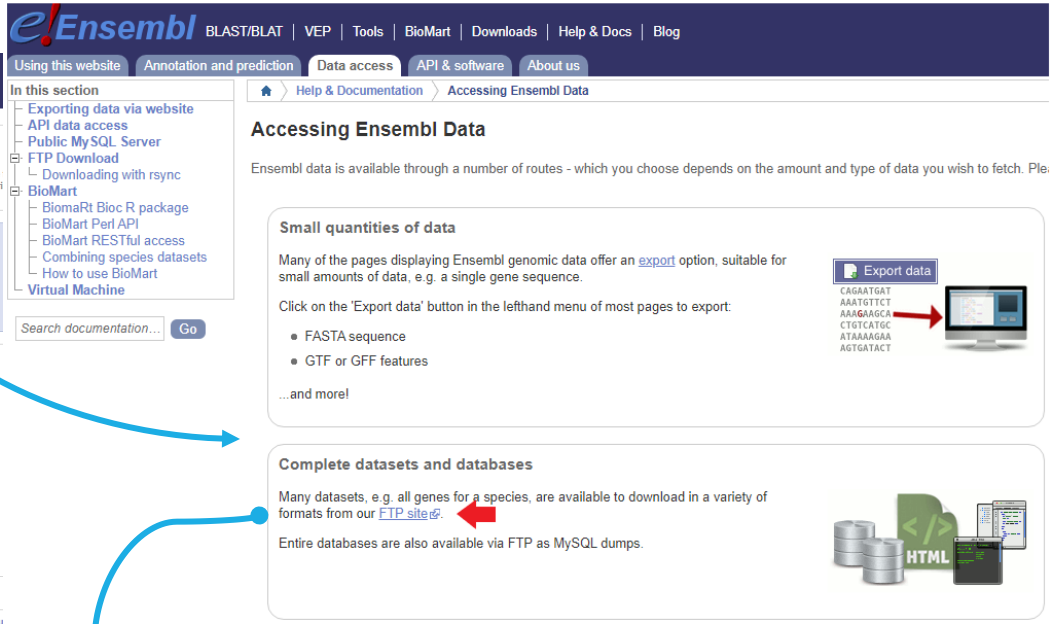
Index of /genomes/all/GCF/000/001/405

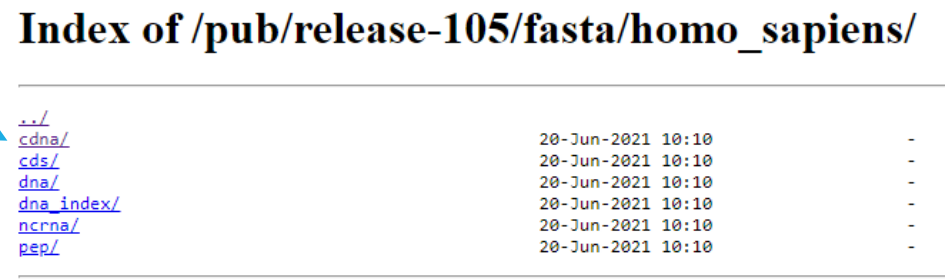
Name	Last modified	Size
Parent Directory		-
GCF_000001405.10_NCBI34/	2020-11-06 20:08	-
GCF_000001405.11_NCBI35/	2020-11-06 20:08	-
GCF_000001405.12_NCBI36/	2020-11-06 20:08	-
GCF_000001405.13_GRCh37/	2020-11-06 20:08	-
GCF_000001405.14_GRCh37.p2/	2020-11-06 20:08	-
GCF_000001405.17_GRCh37.p5/	2020-11-06 20:08	-
GCF_000001405.21_GRCh37.p9/	2020-11-06 20:08	-
GCF_000001405.22_GRCh37.p10/	2020-11-06 20:08	-
GCF_000001405.23_GRCh37.p11/	2020-11-06 20:08	-
GCF_000001405.24_GRCh37.p12/	2020-11-06 20:08	-
GCF_000001405.25_GRCh37.p13/	2022-03-13 15:00	-
GCF_000001405.26_GRCh38/	2020-11-06 20:08	-
GCF_000001405.27_GRCh38.p1/	2020-11-06 20:08	-
GCF_000001405.28_GRCh38.p2/	2020-11-06 20:08	-
GCF_000001405.29_GRCh38.p3/	2020-11-06 20:08	-
GCF_000001405.30_GRCh38.p4/	2020-11-06 20:08	-
GCF_000001405.31_GRCh38.p5/	2021-12-17 15:20	-
GCF_000001405.32_GRCh38.p6/	2020-11-06 20:08	-
GCF_000001405.33_GRCh38.p7/	2020-11-06 20:08	-
GCF_000001405.34_GRCh38.p8/	2020-11-06 20:08	-
GCF_000001405.35_GRCh38.p9/	2020-11-06 20:08	-
GCF_000001405.36_GRCh38.p10/	2020-11-06 20:08	-
GCF_000001405.37_GRCh38.p11/	2020-11-06 20:08	-
GCF_000001405.38_GRCh38.p12/	2020-11-04 23:46	-
GCF_000001405.39_GRCh38.p13/	2021-11-24 15:45	-
GCF_000001405.8_NCBI33/	2020-11-06 20:08	-
GCF_000001405.9_NCBI34/	2020-11-06 20:08	-

https://ftp.ncbi.nlm.nih.gov/genomes/

Downloading Genomes files from ENSEMBL



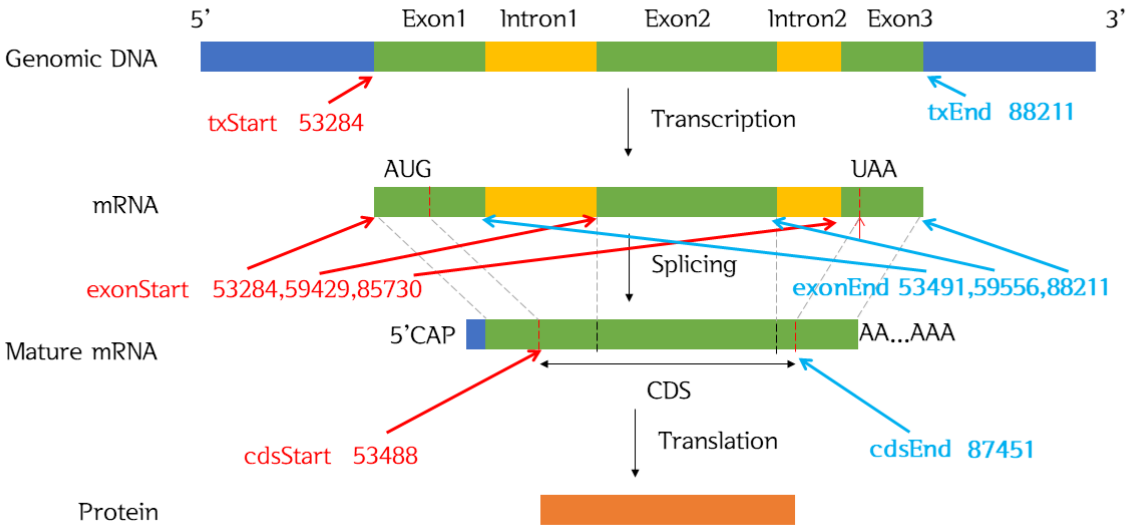




wget or aspera

Reminder of genome structure

.	Chromosome	Strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	Gene
NM_001286052	chr4	+	53284	88211	53488	87451	3	53284,59429,85730,	53491,59556,88211,	ZNF595



Gene

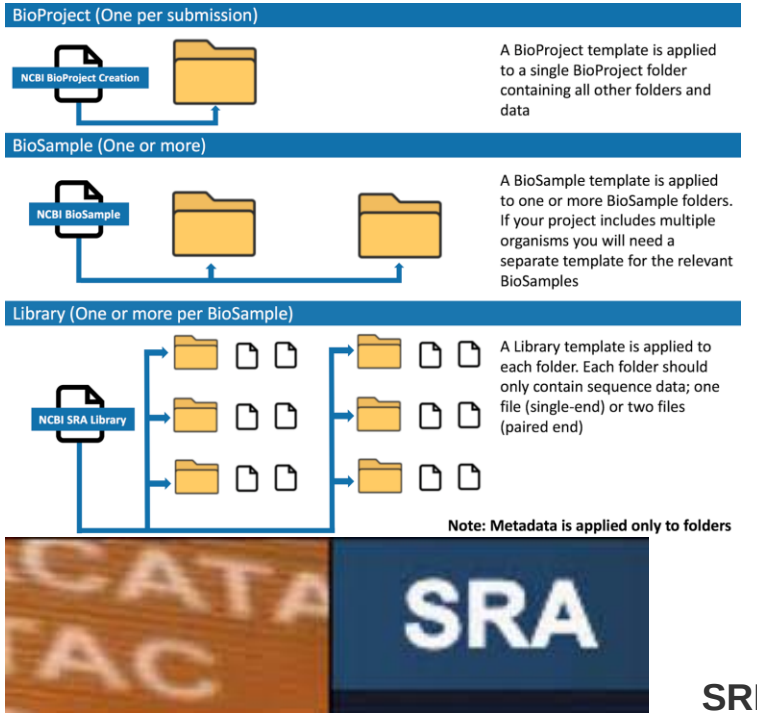
mRNA = Exons + Introns

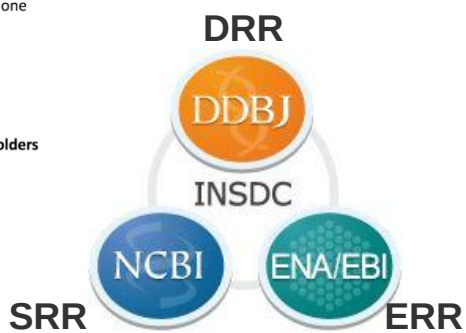
Mature mRNA = Exons (CDS + UTRs)

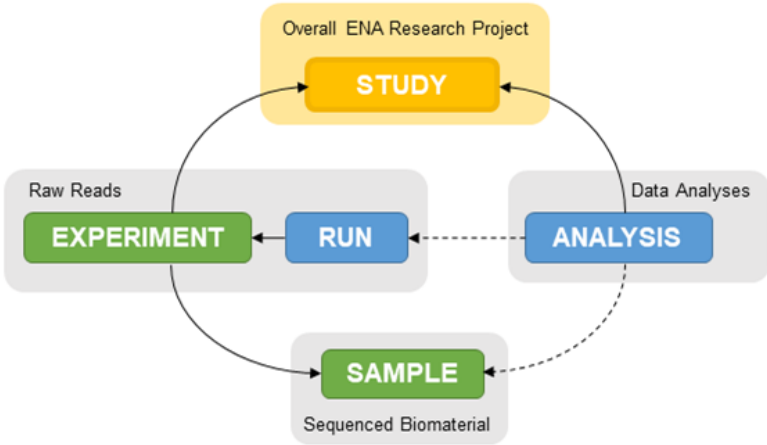
Proteins

wget or aspera

Downloading sequencing files from DBs

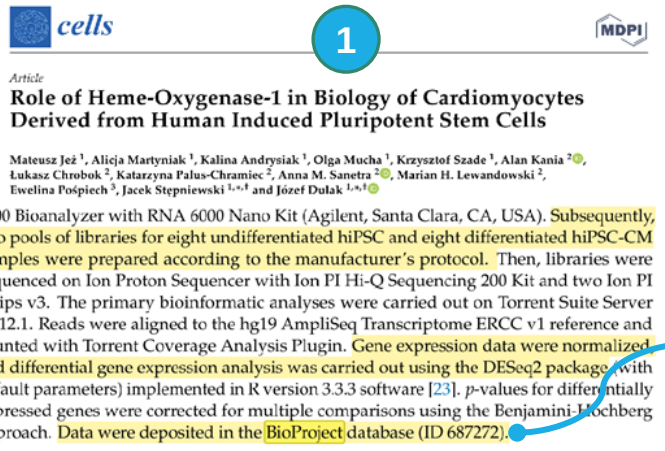






Sequencing data submitted to different databases are stored under SRR to NCBI, ERR to EBI and DRR to DDBJ followed by a number and the same data can be found in all of the three DBs

Downloading Sequence Read Archive (SRA) files from NCBI



cells

MDPI

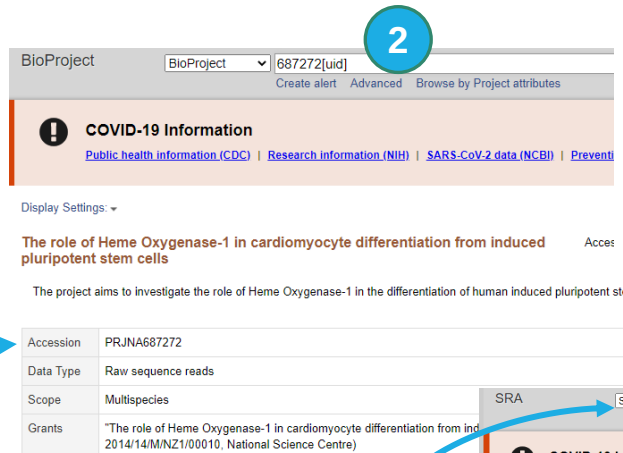
1

Article

Role of Heme-Oxygenase-1 in Biology of Cardiomyocytes Derived from Human Induced Pluripotent Stem Cells

Mateusz Jeż [1], Alicja Martyniak [1], Kalina Andrysiak [1], Olga Mucha [1], Krzysztof Szade [1], Alan Kania [2], Łukasz Chrobok [2], Katarzyna Palus-Chramiec [2], Anna M. Sanetra [2], Marian H. Lewandowski [2], Ewelina Pośpiech [3], Jacek Stępniewski [1,*,†] and Józef Dulak [1,*,†]

2100 Bioanalyzer with RNA 6000 Nano Kit (Agilent, Santa Clara, CA, USA). Subsequently, two pools of libraries for eight undifferentiated hiPSC and eight differentiated hiPSC-CM samples were prepared according to the manufacturer's protocol. Then, libraries were sequenced on Ion Proton Sequencer with Ion PI Hi-Q Sequencing 200 Kit and two Ion PI Chips v3. The primary bioinformatic analyses were carried out on Torrent Suite Server v5.12.1. Reads were aligned to the hg19 AmpliSeq Transcriptome ERCC v1 reference and counted with Torrent Coverage Analysis Plugin. Gene expression data were normalized and differential gene expression analysis was carried out using the DESeq2 package (with default parameters) implemented in R version 3.3.3 software [23]. *p*-values for differentially expressed genes were corrected for multiple comparisons using the Benjamini-Hochberg approach. Data were deposited in the BioProject database (ID 687272).



BioProject

BioProject | 687272[uid]

Create alert | Advanced | Browse by Project attributes

2

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Preventi

Display Settings:

The role of Heme Oxygenase-1 in cardiomyocyte differentiation from induced pluripotent stem cells

Acces

The project aims to investigate the role of Heme Oxygenase-1 in the differentiation of human induced pluripotent stem cells to cardiomyocytes

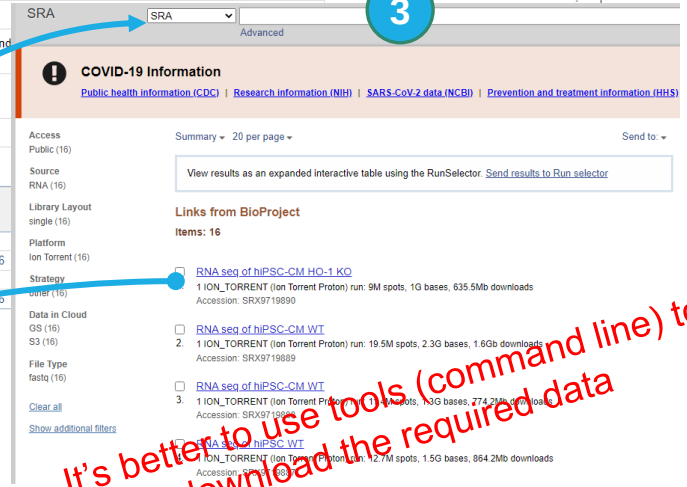
Accession	PRJNA687272
Data Type	Raw sequence reads
Scope	Multispecies
Grants	"The role of Heme Oxygenase-1 in cardiomyocyte differentiation from ind... 2014/14/M/NZ1/00010, National Science Centre)
Submission	Registration date: 22-Dec-2020 **Jagiellonian University**
Relevance	Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	16
OTHER DATASETS	
BioSample	16

SRA Data Details

Parameter	Value
Data volume, Gbases	22
Data volume, Mbytes	14216



SRA

SRA

Advanced

3

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

Access: Public (16)
Source: RNA (16)
Library Layout: single (16)
Platform: Ion Torrent (16)
Strategy: other (16)
Data in Cloud: GS (16), S3 (16)
File Type: fastq (16)
Clear all
Show additional filters

Summary | 20 per page | Send to:

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Links from BioProject

Items: 16

1. RNA seq of hiPSC-CM HO-1 KO
   1 ION_TORRENT (Ion Torrent Proton) run: 9M spots, 1G bases, 635.5Mb downloads
   Accession: SRX9719890
2. RNA seq of hiPSC-CM WT
   1 ION_TORRENT (Ion Torrent Proton) run: 19.5M spots, 2.3G bases, 1.6Gb downloads
   Accession: SRX9719889
3. RNA seq of hiPSC-CM WT
   1 ION_TORRENT (Ion Torrent Proton) run: [illegible] spots, [illegible]3G bases, 274.2Mb downloads
   Accession: SRX9719[illegible]
4. RNA seq of hiPSC WT
   1 ION_TORRENT (Ion Torrent Proton) run: 12.7M spots, 1.5G bases, 864.2Mb downloads
   Accession: SRX9[illegible]

4

SRX9719890: **RNA seq of hiPSC-CM HO-1 KO**
1 ION_TORRENT (Ion Torrent Proton) run: 9M spots, 1G bases, 635.5Mb downloads

Design: Ion AmpliSeq Transcriptome Human Gene Expression Panel

Submitted by: Jagiellonian University

Study: The role of Heme Oxygenase-1 in cardiomyocyte differentiation from induced pluripotent stem cells
PRJNA687272 • SRP298971 • All experiments • All runs
show Abstract

Sample: hiPSC-CM HO-1 KO1
SAMN17140147 • SRS7913872 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: hiPSC-CM HO-1 KO1
Instrument: Ion Torrent Proton
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: RT-PCR
Layout: SINGLE

Runs: 1 run, 9M spots, 1G bases, 635.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR13290858	8,954,125	1G	635.5Mb	2020-12-23

NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries/install scripts of February 10, 2022, version 3.0.0:

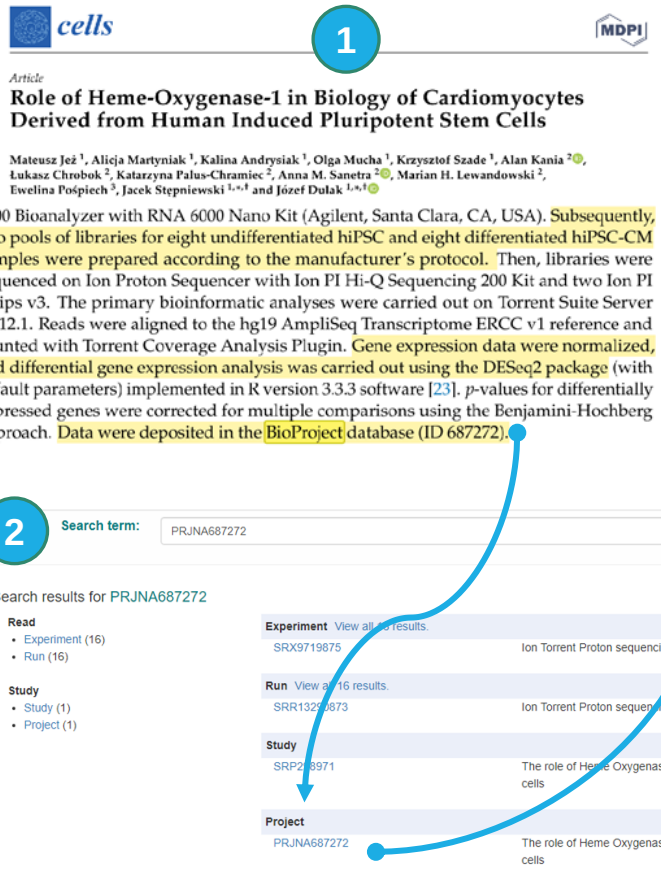
- CentOS Linux 64 bit architecture - non-sudo tar archive
- Ubuntu Linux 64 bit architecture - non-sudo tar archive
- Cloud - apt-get install script - for Debian and Ubuntu - requires sudo permissions
- Cloud - yum install script - for for CentOS - requires sudo permissions
- MacOS 64 bit architecture
- MS Windows 64 bit architecture
- Docker image repository
- md5 checksums

It's better to use tools (command line) to download the required data

Fastq-dump or fasterq-dump

Downloading Sequence Read Archive (SRA) files from NCBI

EMBL-EBI



cells

MDPI

1

Article

Role of Heme-Oxygenase-1 in Biology of Cardiomyocytes Derived from Human Induced Pluripotent Stem Cells

Mateusz Jeż [1], Alicja Martyniak [1], Kalina Andrysiak [1], Olga Mucha [1], Krzysztof Szade [1], Alan Kania [2], Łukasz Chrobok [2], Katarzyna Palus-Chramiec [2], Anna M. Sanetra [2], Marian H. Lewandowski [2], Ewelina Pośpiech [3], Jacek Stępniewski [1,*,†] and Józef Dulak [1,*,†]

2100 Bioanalyzer with RNA 6000 Nano Kit (Agilent, Santa Clara, CA, USA). Subsequently, two pools of libraries for eight undifferentiated hiPSC and eight differentiated hiPSC-CM samples were prepared according to the manufacturer's protocol. Then, libraries were sequenced on Ion Proton Sequencer with Ion PI Hi-Q Sequencing 200 Kit and two Ion PI Chips v3. The primary bioinformatic analyses were carried out on Torrent Suite Server v5.12.1. Reads were aligned to the hg19 AmpliSeq Transcriptome ERCC v1 reference and counted with Torrent Coverage Analysis Plugin. Gene expression data were normalized, and differential gene expression analysis was carried out using the DESeq2 package (with default parameters) implemented in R version 3.3.3 software [23]. *p*-values for differentially expressed genes were corrected for multiple comparisons using the Benjamini-Hochberg approach. Data were deposited in the BioProject database (ID 687272).

2

Search term: PRJNA687272 Search

Search results for PRJNA687272

- **Read**
  - Experiment (16)
  - Run (16)
- **Study**
  - Study (1)
  - Project (1)

Experiment View all 16 results.

SRX9719875 Ion Torrent Proton sequencing; RNA seq of hiPSC-CM HO-1 KO

Run View all 16 results.

SRR13290873 Ion Torrent Proton sequencing; RNA seq of hiPSC-CM HO-1 KO

Study

SRP298971 The role of Heme Oxygenase-1 in cardiomyocyte differentiation from induced pluripotent stem cells

Project

PRJNA687272 The role of Heme Oxygenase-1 in cardiomyocyte differentiation from induced pluripotent stem cells

3

Project: PRJNA687272

The project aims to investigate the role of Heme Oxygenase-1 in the differentiation of human induced pluripotent stem cells to cardiomyocytes

Secondary Study Accession:	SRP298971
Study Title:	The role of Heme Oxygenase-1 in cardiomyocyte differentiation from induced pluripotent stem cells
Center Name:	Jagiellonian University
ENA-REFSEQ:	N
PROJECT-ID:	687272
ENA-FIRST-PUBLIC:	2020-12-24
ENA-LAST-UPDATE:	2020-12-24

Show More

It's better to use tools (command line) to download the required data

Read Files

Show Column Selection

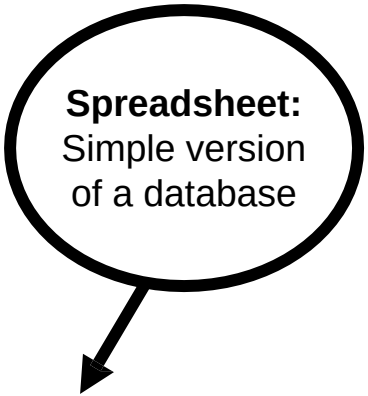
Download report: JSON TSV Download Files as ZIP Download selected files

Download All

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Generated FASTQ files: FTP
PRJNA687272	SAMN17140147	SRX9719890	SRR13290858	9606	Homo sapiens	SRR13290858.fastq.gz
PRJNA687272	SAMN17140146	SRX9719889	SRR13290859	9606	Homo sapiens	SRR13290859.fastq.gz
PRJNA687272	SAMN17140145	SRX9719888	SRR13290860	9606	Homo sapiens	SRR13290860.fastq.gz

Wget or aspera app

Let's Recall the Basics: records



- A collection of records.
- Each record has multiple fields.
- Each field contains specific information.
- Each field contains data of a certain type.
  - *Ex: money, text, integers, dates, addresses*
- Each record has a primary key. A unique identifier that defines the record unambiguously.

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

Let's Recall the Basics: Primary Key

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

gi = Genbank Identifier: **Clave única : Clave primaria**

Cambia con cada actualización del registro correspondiente a la secuencia

Accession Number: **Clave secundaria**

Refiere al mismo locus y secuencia, a pesar de los cambios en la secuencia.

Accession + Version es equivalente al **gi** (representa un identificador único)

Ejemplo: AF405321.2 **Accession**: AF405321 **Version:** 2

Databases: relational databases

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

Relational database:

Normalize a database: distribute repeated sub-elements in several tables, related through a unique identifier (primary key).

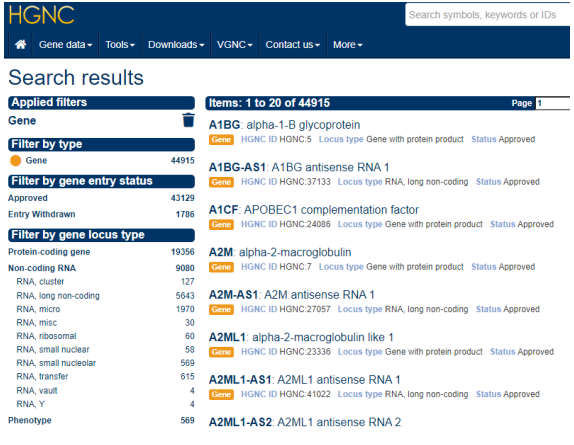
gi	Accession	version	date	Genbank Division	taxid
6226959	NM_000014	3	01/06/2000	PRI	9606
6226762	NM_000014	2	12/10/1999	PRI	9606
4557224	NM_000014	1	04/02/1999	PRI	9606
41	X63129	1	06/06/1996	MAM	9913

taxid	organims	Number of Chromosomes		
9606	homo sapiens	22 diploid + X+Y		
9913	bos taurus	29+X+Y		

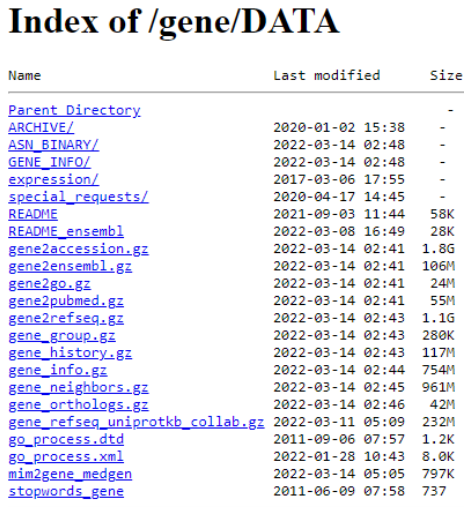
How all the databases are related

How can I identify matches between RefSeq and Ensembl annotation?

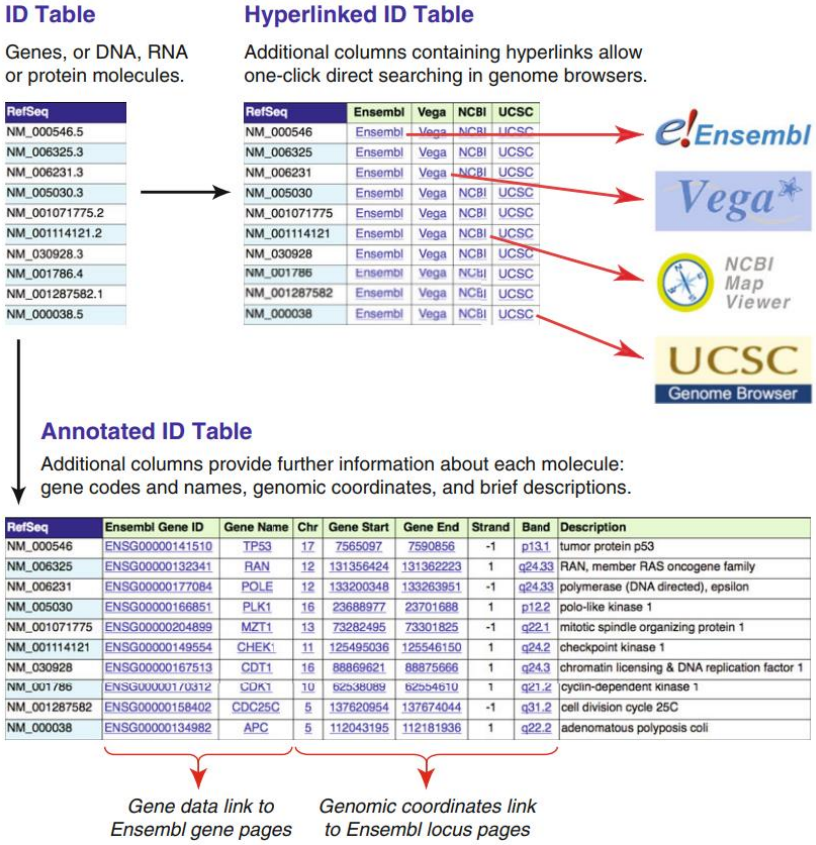
Matches between NCBI and Ensembl annotations can be found in several ways using data provided in Gene, including: the Reference Sequences section of the Full Report display; by using the Gene "matches Ensembl" index property in a query; and in the gene2ensembl FTP file.



https://www.genenames.org



https://ftp.ncbi.nih.gov/gene/DATA/



Protein Databases

Protein Databases

Since most of the sequences in protein DBs come from machine translation of nucleotide sequences, these are usually the first secondary databases.

GenePept:

is the protein database derived from machine translations of the INSDC CDS. (It does not contain sequenced proteins as such!). It is redundant.

trEMBL:

Automatic translation of the EMBL. It is automatically annotated and is not reviewed.

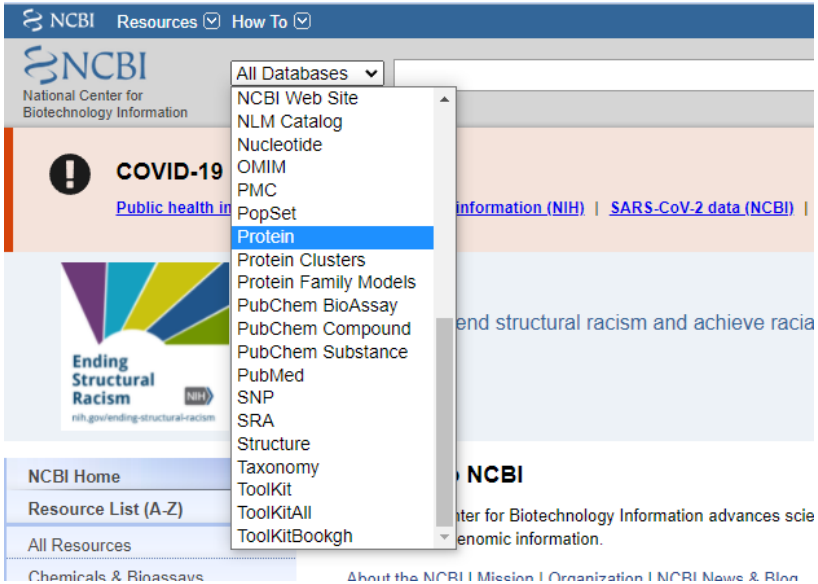
Protein:

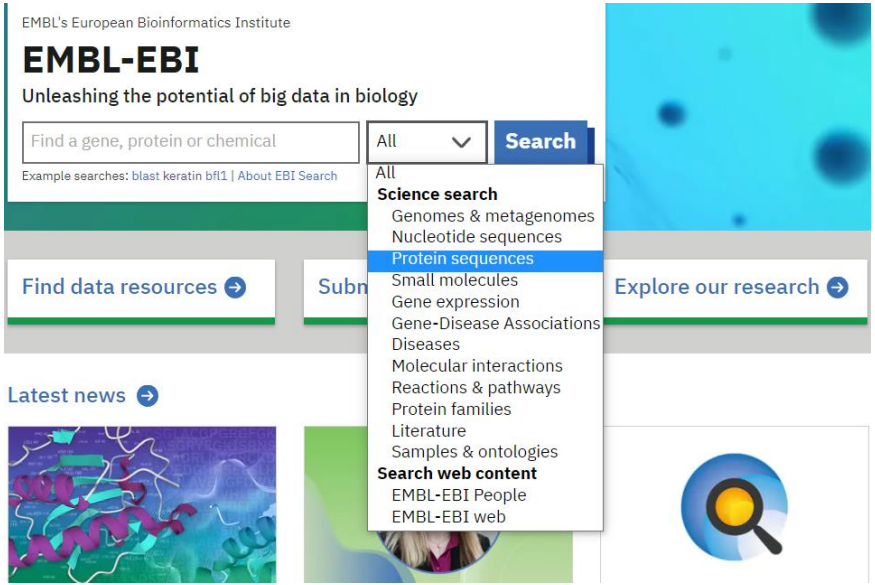
NCBI ENTREZ Protein Database compiling GenPept + RefSeq + Swiss-Prot + PIR + RPF + PDB records

UniProt:

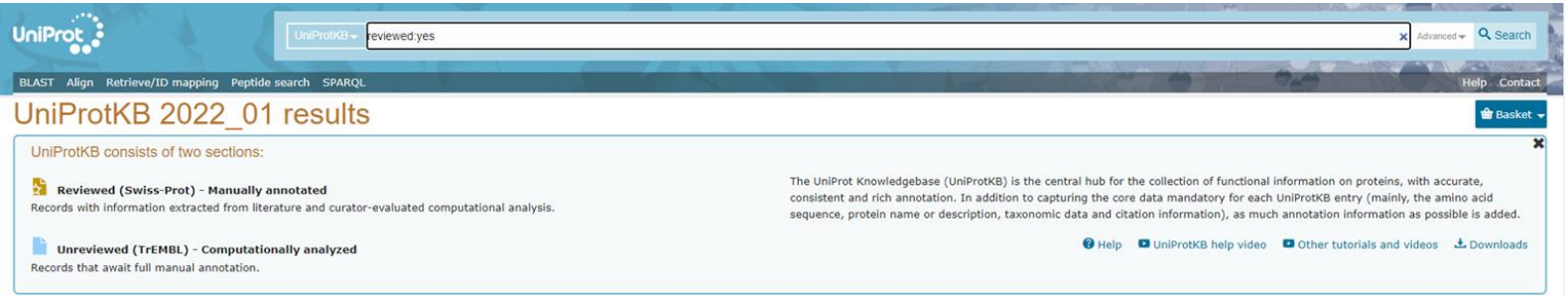
It is perhaps the quintessential protein database. It contains the sequences derived from SwPr, TrEMBL and PIR. It is manually annotated and reviewed.

Protein Databases





UniProt Knowledgebase



- Created in 1986 by Amos Bairoch (EBI)
- Curated Database
- It has high level annotations
- – Function, post-translational modifications, variants, organization and structure of domains
- It is the source of most 2ria databases
- Little redundancy or no redundancy
- They are more reliable

The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference.

UniProtKB comprises two sections:

UniProtKB/Swiss-Prot which is manually annotated and is reviewed and **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed.

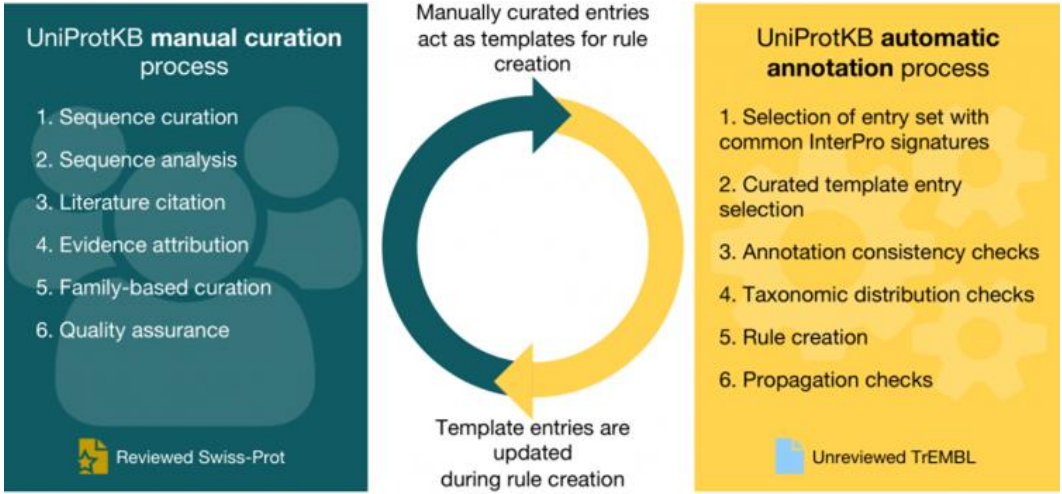
Swiss-Prot:
http://www.expasy.ch/sprot/

UniProt Knowledgebase

UniProtKB: The UniProt Knowledgebase (UniProtKB), the centrepiece of the UniProt Consortium's activities, is an expertly and richly curated protein database, consisting of two sections called UniProtKB/Swiss-Prot and UniProtKB/TrEMBL:

1. **UniProtKB/Swiss-Prot:** UniProtKB/Swiss-Prot contains high-quality expertly curated and non-redundant protein sequence records. Expert curation consists of a critical review of experimental and predicted data for each protein by a team of biologists, as well as manual verification of each protein sequence. UniProt curators extract biological information from the literature and perform numerous computational analyses. UniProtKB/Swiss-Prot aims to provide all known relevant information about a particular protein. Data captured from the scientific literature includes information on protein and gene names, function, catalytic activity, cofactors, subcellular location, protein-protein interactions and much more.

2. **UniProtKB/TrEMBL:** contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot. The database is enriched with automated classification and annotation.



Uniprot: sections and redundancy

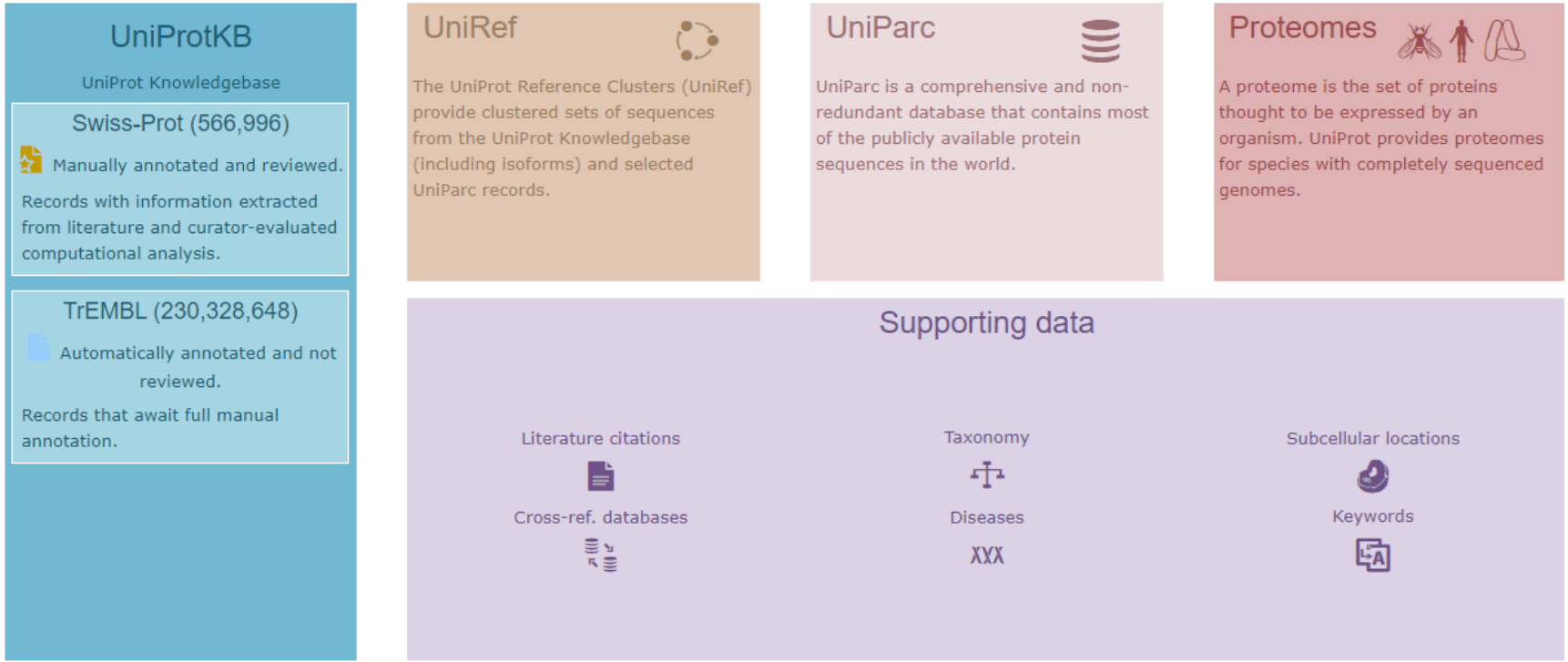
UniParc:

- Where the new sequences from the different DBs are loaded
- It is NOT redundant as each protein is assigned 1 unique ID (and cross references to the multiple sources).
- Sequences equal throughout their length correspond to the same ID, regardless of the species to which they belong.
- UniProt Knowledgebase:
- contains as many annotations as possible about the biological information.
- Annotations are divided into Manual (SwPr section) and Automatic (TrEMBL).
- The redundancy criterion is 1 record 1 gene (different from UniParc)

UniRef:

- Non-redundant secondary database x comparison and grouping.
- Group proteins by sequence identity.
- Uniref100, each record groups the proteins with 100% identity,
- Uniref90 all prots with 90% identity
- Uniref50….

Uniprot: sections and redundancy



Uniprot: sections and redundancy

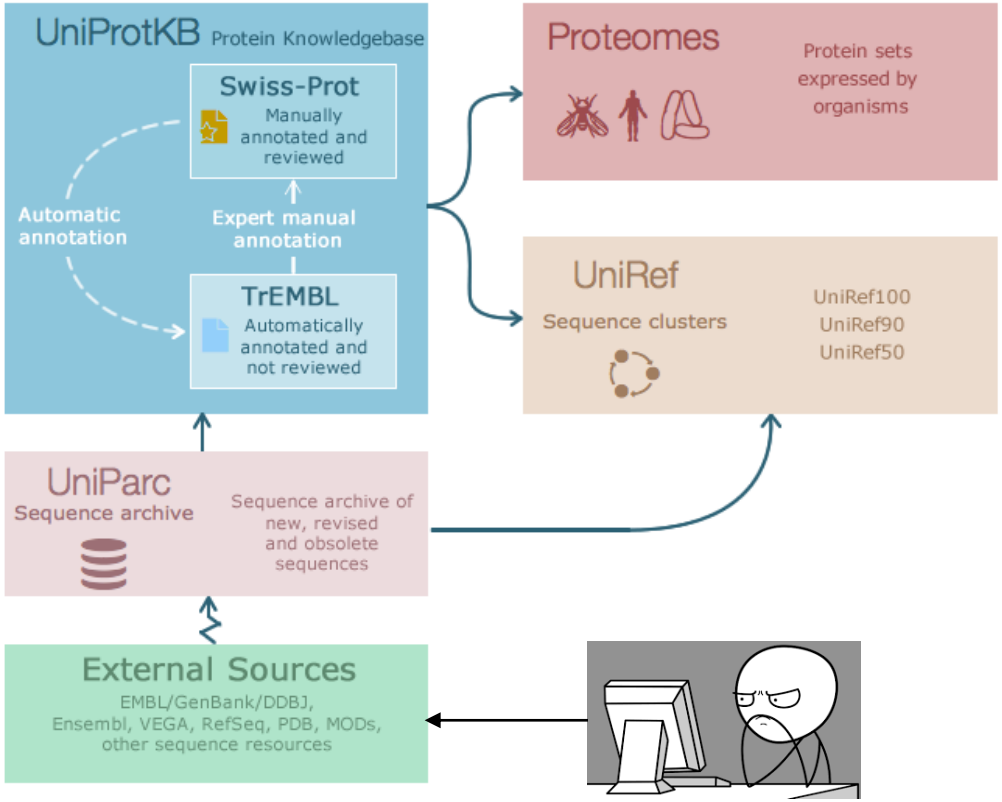
4 components of UniProt

UniParc
- Complete history of sequences *(no annotation)*
- Cross-links to external sequence sources

UniProtKB
- *Swiss-Prot*: non-redundant, manual annotation
- *TrEMBL*: redundant, automatic annotation

UniMES
- Sequences from metagenomic projects

UniRef
- Combines sequences (speed searching)
- UniRef100, UniRef90, UniRef50



Uniprot

•Sequence sources

•The default raw sequence data for UniProtKB are:

- DDBJ/ENA/GenBank coding sequence (CDS) translations;
- sequences of PDB structures;
- sequences from Ensembl and RefSeq;
- data derived from amino acid sequences that are directly submitted to UniProtKB or scanned from the literature.

•What is not included

- The following protein sequence types are not included in UniProtKB but are stored in the UniProt Archive (UniParc):
  - small fragments;
  - synthetic sequences;
  - most non-germline immunoglobulins and T-cell receptors;
  - most patent sequences;
  - pseudogenes;
  - sequences from redundant proteomes;

Downloads from Uniprot

UniProtKB

Parent directory

- Reviewed (Swiss-Prot)[1] / FAQ
- Unreviewed (TrEMBL)[1] / FAQ
- Isoform sequences[1] / FAQ
- Taxonomic divisions / README
- Reference proteomes / README
- Pan proteomes / README
- ID mapping / README
- Proteomics mapping / README
- Variants / README
- Genome annotation tracks / README
- Documents
- XML schema

UniRef[1]

Parent directory

- UniRef100 / README
- UniRef90 / README
- UniRef50 / README
- XML schema

UniParc[1]

Parent directory / README

- UniParc (Sequence Archive) / Help
- XML schema

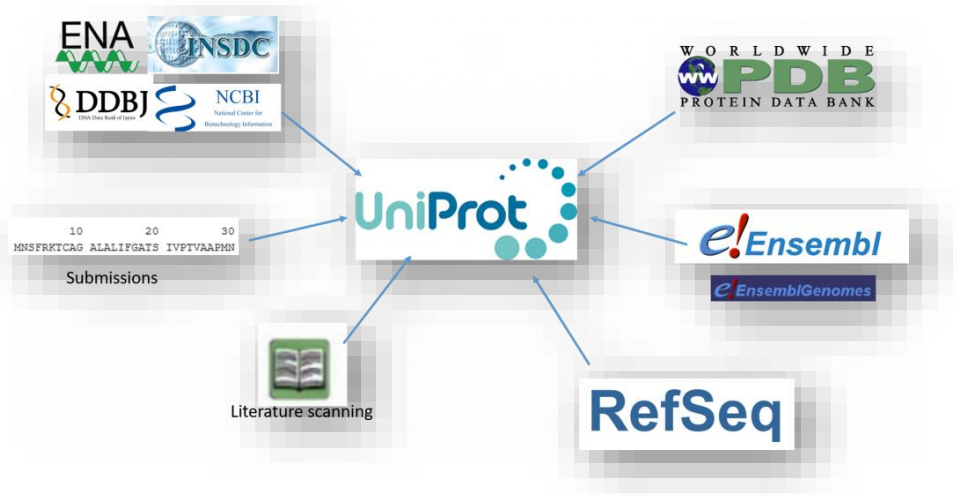
UniProt RDF distribution[1]

Parent directory / README

- OWL schema

UniProt information

- Releases notes
- Previous releases



Index of /pub/databases/uniprot

Name	Last modified	Size	Description
Parent Directory		-	
LICENSE	2021-12-13 11:44	384	
README	2021-11-17 10:00	3.9K	
current_release/	2022-03-02 10:00	-	
knowledgebase/	2022-03-02 10:00	-	
pre_release/	2021-12-13 13:38	-	
previous_major_releases/	2022-03-01 10:10	-	
previous_releases/	2022-03-01 10:10	-	
relnotes.txt	2022-03-02 10:00	1.0K	
uniref/	2022-03-02 10:00	-	

Index of /pub/databases/uniprot/current_release/knowledgebase/idmapping

Name	Last modified	Size	Description
Parent Directory		-	
README	2022-03-02 10:00	3.7K	
RELEASE.metalink	2022-03-02 10:00	5.3K	
by_organism/	2022-03-02 10:00	-	
idmapping.dat.2015_03.gz	2022-03-02 10:00	6.4G	
idmapping.dat.example	2022-03-02 10:00	258K	
idmapping.dat.gz	2022-03-02 10:00	21G	
idmapping_selected.tab.2015_03.gz	2022-03-02 10:00	4.1G	
idmapping_selected.tab.example	2022-03-02 10:00	328K	
idmapping_selected.tab.gz	2022-03-02 10:00	11G	

ftp source Uniprot
https://ftp.uniprot.org/pub/databases/uniprot/
Webserver Uniprot
https://www.uniprot.org/downloads#uniprotkblink

Secondary DBs of proteins

“Patterns

Secondary structures or domains. They vary according to the source of the proteins and the analysis that is carried out on them.

BLOCKS: PROSITE/PRINTS Aligned Motifs

PROSITE: Regular expressions over Swiss-prot

PRINTS: Set of motifs that define a family on Swiss-prot/TrEMBL

PFAM: Markov models on Swiss-prot (HMM)

INTERPRO: Integrates information from many domain databases.

Secondary DBs of proteins

“Patterns

Three-dimensional structures of macromolecules with the space coordinates of each atom.

PDB: Three-Dimensional Structures Master Database

CATH: Classification of PDBs in different functional and structural groups

MMDB: subset of PDB maintained by NCBI

MSD: subset of PDB maintained by EBI

Secondary DBs of proteins

“Structures”

BRENDA is an enzyme information system representing one of the most comprehensive enzyme repositories.

BRENDA contains:

- Comprises **molecular and biochemical information on enzymes**
- Every classified enzyme is characterized with respect to its biochemical reaction.
- Kinetic properties of the substrates and products are described in detail.
- **enzyme-specific data** manually extracted from primary scientific literature and additional data derived from automatic information retrieval methods such as text mining.
- more than 4800 EC numbers that are classified according to the IUBMB.

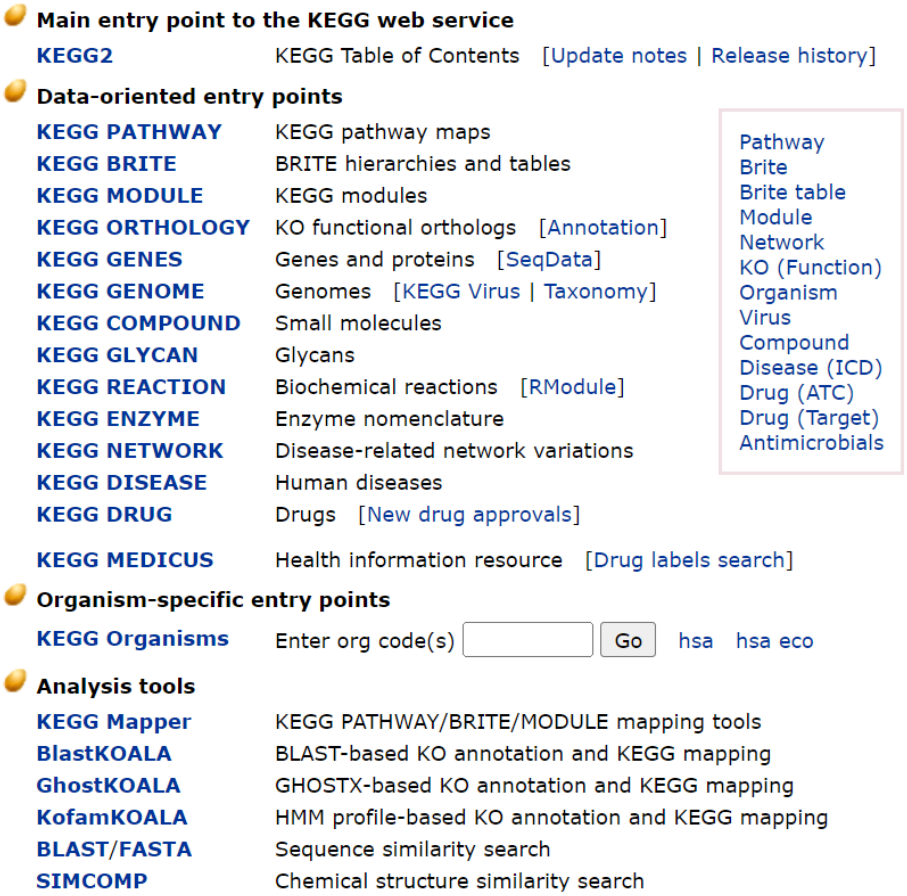
Data fields cover information on:

- the enzyme's nomenclature
- reactionand specificity
- enzyme structure
- isolation and preparation
- enzyme stability
- kineticparameters such as Km value and turnover number
- occurrence and localization
- mutants and engineered enzymes
- application of enzymes
- ligand-related data

BRENDA

The Comprehensive Enzyme Information System

Enzyme DAtabase

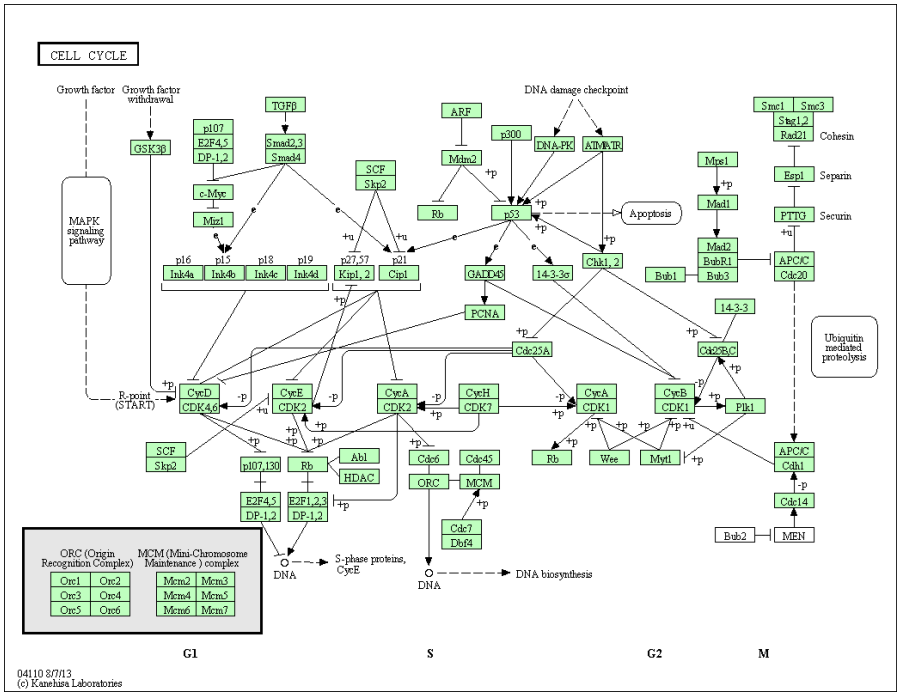


KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-levelfunctions and utilities of the biological system, such asthe cell, the organism and the ecosystem, frommolecular-level information, especially large-scalemolecular datasets generated by genome sequencing andother high-throughput experimental technologies

Reactome & KEGG

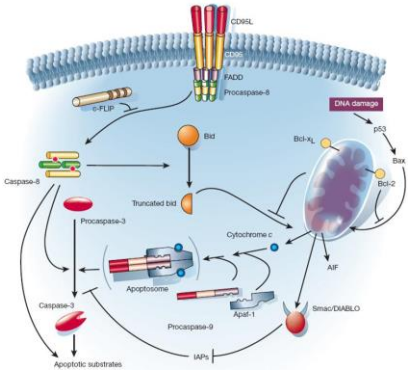
- explicitly describe biological processes as a series of biochemical reactions.
- represents many events and states found in biology.



KEGG: Kyoto Encyclopedia of Genes and Genomes

- KEGG is a collection of biological information compiled from published material (curated database)
- Includes information on genes, proteins, metabolic pathways, molecular interactions, and biochemical reactions associated with specific organisms
- Provides a relationship (map) for how these components are organized in a cellular structure or reaction pathway.

Reactome Rationale



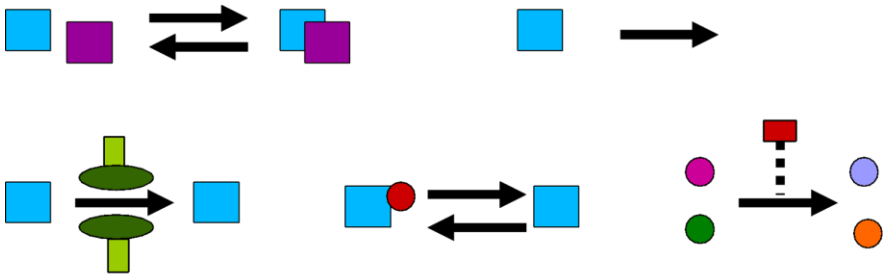
A picture paints a thousand words…

...BUT

- is not computationally accessible
- doesn't convey enough detail

Theory - Reactions

- Basic "unit" of Reactome
- Represents many events and states found in biology.

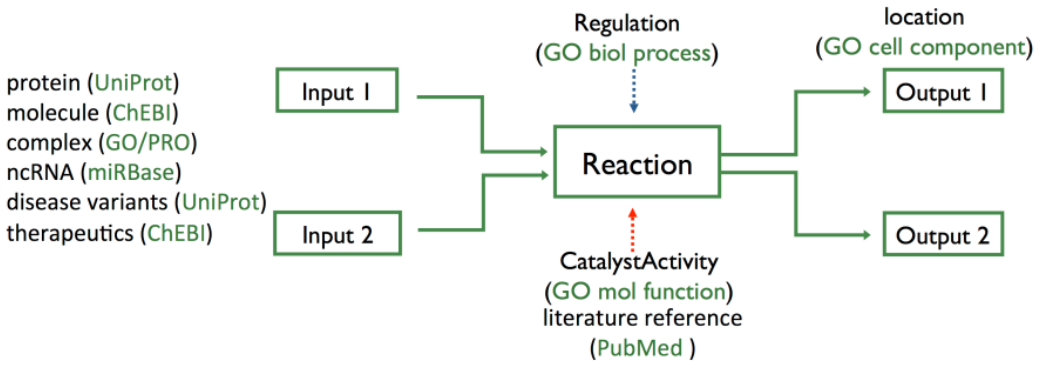


reactome

Reactome is a free, expert-authored, peer-reviewed knowledgebase of pathways and reactions in human biology

- Data Analysis and Visualization Tools
- Data downloads – interaction, BioPAX, SBML, etc.
- Curated human data are used to infer orthologous events in 22 non-human species

Reactome Reaction & Pathway



reactome

Reactome is a free, expert-authored, peer-reviewed knowledgebase of pathways and reactions in human biology

- Pathway databases
  - INOH, PANTHER, KEGG, CellMap, PID
- Bimolecular interactions
  - Human interactions from BioGrid, BIND, IntACT, HRPD
  - Interlogs from yeast, worm, fly
- Shared GO terms
- Gene co-expression data
  - Multiple sources
- Transcription factors/targets
  - TRED
- Literature mining
  - GeneWays

Corona Data Sources & Combined Network

10,956 proteins
209,988 interactions

46% coverage of UniProt

Apoptosis	Axon guidance
Cell-cell adhesion systems	Cell Cycle Checkpoints
DNA Replication	Diabetes pathways
Gene Expression	HIV Infection
Integration of energy metabolism	Integrin cell surface interactions
Metabolism of amino acids	Metabolism of carbohydrates
Metabolism of polyamines	Metabolism of proteins
Metablism of nucleotides	Metabolism of porphyrins
Regulatory RNA pathways	Signaling by BMP
Signaling by GPCR	Signaling by PDGF
Signalling by NGF	Signaling by Notch
Signaling by TGF beta	Signaling by VEGF
Telomere Maintenance	Transcription

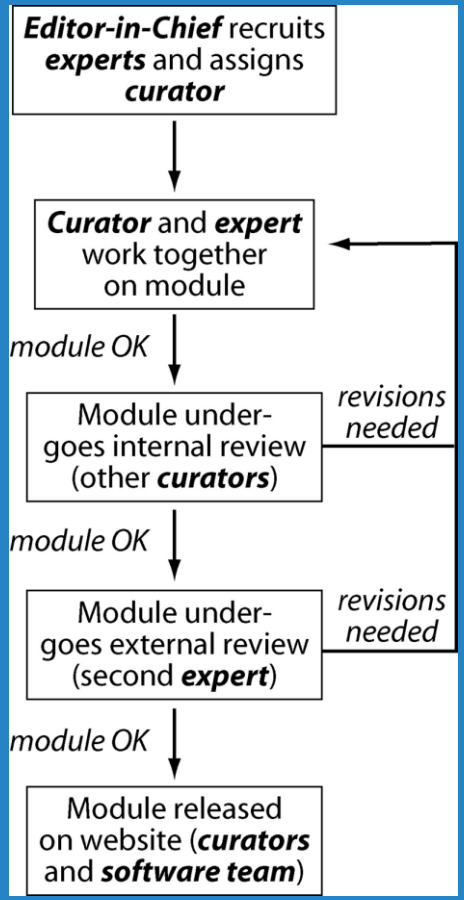
Biological oxidations	Botulinum neurotoxicity
Cell Cycle, Mitotic	DNA Repair
Electron Transport Chain	Gap junction trafficking and regulation
Hemostasis	Influenza Infection
Metabolism of lipids and lipoproteins	Membrane Trafficking
Metabolism of nitric oxide	Metabolism of non-coding RNA
Metabolism of vitamins and cofactors	Muscle contraction
Pyruvate metabolism and TCA cycle	Regulation of beta-cell development
Signaling by EGFR	Signaling by FGFR
Signaling in Immune system	Signaling by Insulin receptor
Opioid Signalling	Signaling by Rho GTPases
Signaling by Wnt	Synaptic Transmission
Transmembrane transport of small molecules	mRNA Processing

3.4.21.4: trypsin

This is an abbreviated version!
For detailed information about trypsin, go to the full flat file.

Word Map on EC 3.4.21.4

reactome

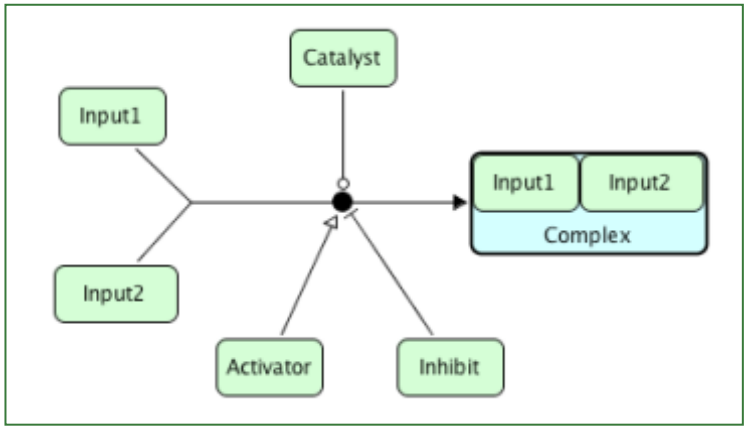


What is a Functional Interaction?

Convert reactions in pathways and into pair-wise relationships

Functional Interaction: an interaction in which two proteins are involved in the same reaction as input, catalyst, activator and/or inhibitor, or as components in a complex

Reaction

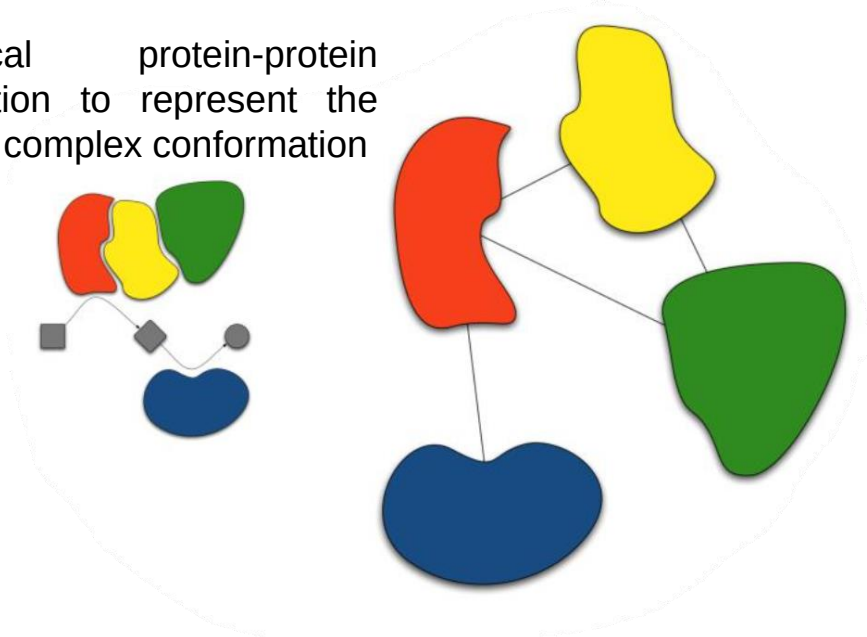


Functional Interactions

Input1-Input2, Input1-Catalyst, Input1-Activator, Input1-Inhibitor, Input2-Catalyst, Input2-Activator, Input2-Inhibtior, Catalyst-Activator, Catalyst-Inhibitor, Activator-Inhibitor

STRING integrates information and predicts interactions • You can always go to the sources • Proteins mode: specific species • COG mode: more coverage, especially for prokaryotic genes

Theorical protein-protein interaction to represent the protein complex conformation



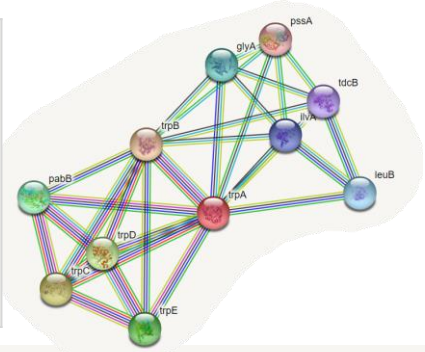
Protein-protein interaction through network interaction network to represent the protein complex conformation

STRING

STRING imports protein association knowledge from databases of physical interaction and databases of curated biological pathway knowledge (MINT, HPRD, BIND, DIP, BioGRID, KEGG, Reactome, IntAct, EcoCyc, NCI-Nature Pathway Interaction Database, GO).

trpA

Tryptophan synthase, alpha subunit; The alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate (268 aa)



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

Network currently showing
Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

Neighborhood
Groups of genes that are frequently observed in each other's genomic neighborhood.

Experiments
Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

Fusion
Genes that are sometimes fused into single open reading frames.

Databases
Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

Cooccurrence
Gene families whose occurrence patterns across genomes show similarities.

Textmining
Automated, unsupervised textmining - searching for proteins that are frequently mentioned together.

Coexpression
Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

STRING

BioGRID Database of Protein, Genetic and Chemical Interactions

Algunas bases de datos biológicas

• Secuencias nucleotídicas (ADN y ARN): la colaboración de las tres bases de datos más importantes hace posible acceder a casi toda la información de secuencias de nucleótidos desde cualquiera de sus tres sedes
Bases de datos de ***EMBL*** en el *European Bioinformatics Institute* (*EMBL-EBI*). Enlace externo base de datos de nucleótidos de EMBL-EBI
DNA Data Bank of Japan (DDJB). Enlace externo DDJB
GenBank en el *National Center for Biological Information (NCBI)*. Enlace externo GenBank
Si bien son mantenidas por distintos organismos en distintos países, existe una coordinación entre las distintas bases. Una secuencia enviada a cualquiera de las bases se verá reflejada en las otras dos en aproximadamente una semana, ya que esa es la frecuencia de actualización entre las distintas bases genéticas. Por este motivo es indistinto que base se use para enviar nuevas secuencias, aunque normalmente los europeos utilizan EMBL y los americanos GenBank.
• Proteínas: bases de datos de secuencias, estructuras, e información relacionada
UniProtKB/Swiss-Prot contiene secuencias anotadas o comentadas, es decir, cada secuencia ha sido revisada, documentada y enlazada a otras bases de datos. Enlaces externos UniProtKB, Swissprot en el EBI ***UniProtKB/TrEMBL*** por *Translation of EMBL Nucleotide Sequence Database* incluye la traducción de todas las secuencias codificantes derivadas del (EMBL) y que todavía no han podido ser anotadas en Swiss-Prot. Enlaces externos TrEMBL, UniProtKB '***PIR*** por *Protein Information Resource* está dividida en cuatro sub-bases que tienen un nivel de anotación decreciente. Enlace externo PIR
'***ENZYME*** enlaza la clasificación de actividades enzimáticas completa a las secuencias de Swiss-Prot. Enlace externo ENZYME
'***PROSITE*** contiene información sobre la estructura secundaria de proteínas, familias, dominios, etc. Enlace externo PROSITE
'***InterPro*** integra la información de diversas bases de datos de estructura secundaria como PROSITE, proporcionando enlaces a otras bases de datos e información más extensa. Enlace externo INTERPRO
'***Protein Data Bank*** ***(PDB)*** es la base de datos de estructura terciaria 3D de proteínas que han sido cristalizadas. Enlace externo PDB
• Expresión
El portal de EMBL-EBI ofrece una variedad de bases de datos de expresión génica. Enlace externo a bases de datos de expresión de EMBL-EBI
• Interactomas, reactomas y rutas metabólicas
Reactome es una base de datos curada y revisada de EMBL-EBI de rutas de interacción y reacción de proteínas y enzimas. Enlace externo a Reactome
APID[6] es una base de datos de interacciones proteína-proteína que incluye interactomas completos para múltiples especies. Enlace externo a APID
• Variación genética (SNPs) y enfermedad
dbSNP de NCBI, ofrece un repositorio central de variaciones genéticas que comprenden sustituciones simples de nucleótidos y polimorfismos de inserciones y deleciones cortas. Enlace a dbSNP
COSMIC es un catálogo de mutaciones somáticas en cáncer, mantenida por el Wellcome Trust Sanger Institute. Enlace externo a COSMIC
'***OMIM*** por *Online Mendelian Inheritance in Man* es un catálogo de genes humanos relacionados con desórdenes genéticos. Enlace externo OMIM
• Literatura
Pubmed da acceso gratuito al índice de publicaciones de la Biblioteca Nacional de Medicina (NLM), con enlaces a artículos completos. Enlace externo PubMed
• Ontología
El proyecto de Ontología Génica (GO) es un esfuerzo colaborativo que surgió de la necesidad de tener descriptores consistentes de los productos de genes depositados en distintas bases de datos. Enlace externo a Gene Ontology Consortium
• genomas
Ensembl integra genomas eucariotas grandes, por el momento contiene genoma humano, ratón, rata, fugu, zebrafish, mosquito, Drosophila, *C. elegans*, y *C. briggsae*. Enlace externo Ensembl
Genomes server y **TIGR** son portales con información o enlaces de todos los genomas secuenciados por el momento, desde virus a humanos. Enlace externo Genome Server, enlace externo TIGR
Wormbase es el portal del genoma de gusano *C. elegans*. Enlace externo Wormbase
Flybase es el portal de la mosca de la fruta *Drosophila melanogaster*. Enlace externo Flybase
• Otras
Taxonomy es el portal de clasificación taxonómica de organismos. Enlace externo Taxonomy Browser
Xenobase es el portal del organismo modelo Xenopus laevis. Enlace externo: Xenbase
TAIR (The Arabidopsis Information Resource) es el portal de la planta modelo *Arabidopsis thaliana.* Enlace externo Arabidopsis
GYPSY, base de datos de elementos genéticos móviles. Enlace externo The GYPSY Database of Mobile Genetic Elements

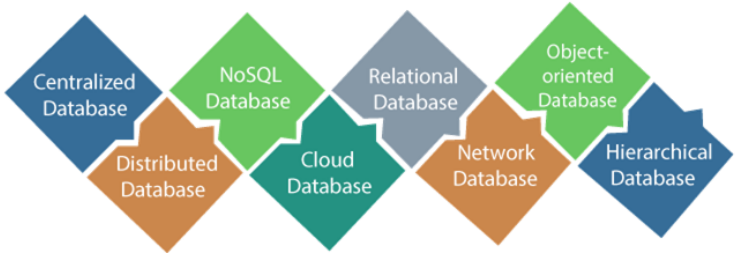
Database architecture

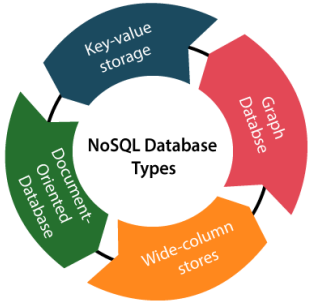
Database architecture

There are several particularly popular ways to organize information in a database.

- classic
- Relational
- object oriented
- unstructured
- Model data to DB Graphs

Type of BioDBs based on their architecture





The **main purpose** of the biological database is to operate a large amount of information by storing, retrieving, and managing data which must be easily retrieved.

Almost all bioDBs are based on relational databases

SQL	NoSQL
• SQL databases are table based databases.	• NoSQL databases are document based, key-value pairs, graph databases.
• Have **predefined schema**.	• Have **dynamic schema**.
• Are vertically scalable.	• Are horizontally scalable.
• Use SQL (Structured Query Language) for defining and manipulating the data.	• Focused on the collection of documents.
	• Not ideal for complex queries.
• A good fit for the complex query intensive environment	• Follow the **Brewers CAP theorem** (Consistency, Availability and Partition tolerance)
• Emphasize **ACID** properties (Atomicity, Consistency, Isolation and Durability)	• Examples include: MongoDB, BigTable, Redis, RavenDb, Cassandra, Hbase, Neo4j and CouchDb.
• Examples include: MySql, Oracle, Sqlite, Postgres and MS-SQL.	

Type of BioDBs based on their architecture

Relational Databases and type of data

A Relational database is based on the relational data model, which stores data in the form of rows(tuple) and columns(attributes), and together forms a table(relation).

- Relational database model has two main terminologies called instance and schema:
  - The **instance** is a table with rows or columns
  - **Schema** specifies the structure like name of the relation, type of each column and name.
- Uses SQL for storing, manipulating, as well as maintaining the data.
- The collection of data items have pre-defined relationships between them.
- The format of the data has to be predefined
- Stores and provides access to data points that are related to one another.

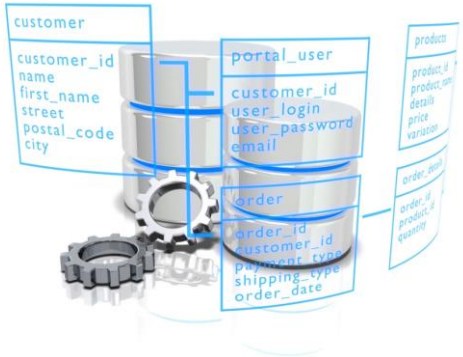
Each field in a database contains a particular type of data.
211203
It's a number?, Is it text?, Is it a date?

Numeric (integers, decimals)
Text
Dates (DD/MM/YYYY, HH:MM:SS)
Logical (boolean) = true / false
Geometric (point, line, circle, polygons, etc.)

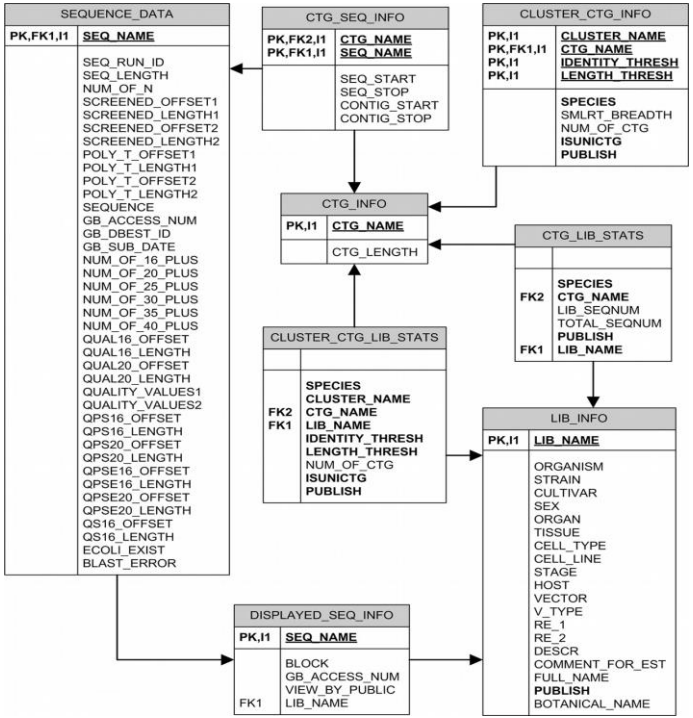
Type of BioDBs based on their architecture

Relational Databases and type of data

The distribution of data in fields within a table and of the relationships between tables and their fields is what is called the layout or schema. It is the list of tables of a database with a complete description of the columns of each table, the types of data that each column contains and the relationships between tables.



Each table has a column where the data is repeated but is the code that relates each table with each other



Type of BioDBs based on their architecture

Relational Databases and type of data

Example of relations:

Proteins ↔ Bibliographic references

Accession	Description	MW	pI
AF1234	Malate dehydrogenase	36000	6.4
AM44432	Cysteine proteinase	45000	4.5

Accession	PubMed ID
AF1234	1234556
AF1234	23445

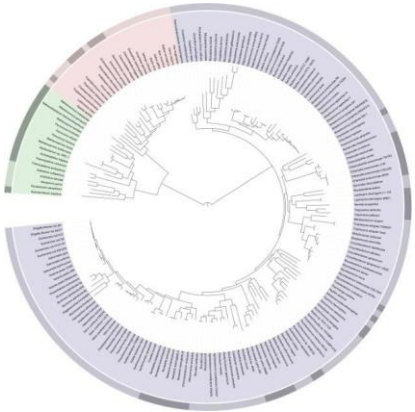
PubMed ID	Journal	Year	Title	Vol
1234556	J Biol Chem	1978	The malate dehydrogenase ...	5
23445	Biochem J	1982	A malate dehydrogenase from ...	13

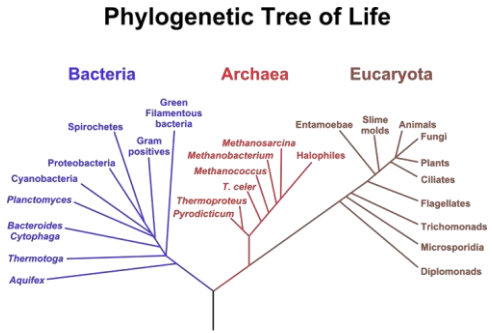
Each table has a column where the data is repeated but is the code that relates each table with each other

Representations of trees and graphs

Example: representation in relational form of trees and graphs

- **Hierarchically** structured information
- Taxonomy (NCBI), SCOP (Structural Classification of Proteins)





Relational modeling of biological data: trees and graphs. Aaron J. Mackey. http://www.oreillynet.com/pub/a/network/2002/11/27/bioconf.html

Adjacency list: Queries

- **What queries can we make on the data organized in the form of an 'adjacency list'?**
  - We can find the immediately superior taxon of any taxonomic element.
  - We can find terminal taxa without 'children'
  - We can find a taxon (or taxa) by searching for them by name
- **And which ones are difficult to do with this representation of the data?**
  - **Can we find all the 'child' taxa of a given taxon?**
    - Typical examples of this type of queries: search for all mammals, all vertebrates, or all members of the order Apicomplexa.
    - How would you do this query? Is it possible to answer these questions with a single database query? How many queries should they do?

Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostome; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Primates; Catarrhini; Hominidae; Homo/Pan/Gorilla Group; Homo; Homo sapiens

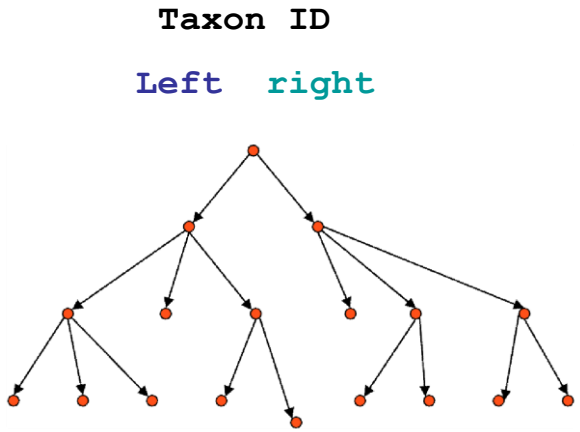
Representación relacional de árboles: nested set

	Campo	Tipo
PK	Taxon_id	entero
FK	Parent_id	entero
	Left_id	entero
	Right_id	entero
	Nombre	texto

Taxon	Nombre	Parent	Left	Right
1	Root	NULL	1	323458
2	Bacteria	1	21703	87862
3	Archaea	1	87863	92266
4	Eukaryota	1	92267	323456
1224	Proteobacteria	2	23982	49591
...	...	...	...	...
543	Enterobacteriaceae	1236	26681	27938
561	Escherichia	543	26852	26891
562	Escherichia coli	561	26853	26868
83333	Escherichia coli K12	562	26856	26857

There are different ways of moving forward

- Depth-first
- Breadth-first



The left and right values are arbitrary numbers, but must satisfy the following property:

"For each 'parent-child' pair the values of the child have to be within the values of the parent"

Object-oriented Databases

Object-oriented modeling

- attempts to simplify the handling of complex information by increasing the level of abstraction
- It does not deal with “data” but with “objects”

An object

- It has descriptive properties
- Interact with other objects

It allows building complex models that include the interactions between objects.

Models can be used to derive answers that are not in the database.

Graph-oriented Databases

- Many companies have data that is of little use because it is not structured, they do not know the relationship between them.
- Graph-oriented databases (BDOG) help to find relationships and make sense of the whole puzzle.
- It must be absolutely normalized, this means that each table would have a single column and each relation only two, with this it is achieved that any change in the structure of the information has only a local effect.
- **Graph-oriented databases represent information as nodes of a graph and their relationships with its edges**, so that graph theory can be used to traverse the database since it can describe attributes of the nodes (entities) and the edges (relationships).