

Public Databases in Health and Life Sciences

“the potential to translate big data into big discovery”



Academic year 2023-2024

Public Databases in Health and Life Sciences



Lesson 1. Organizing biological knowledge in databases.

- Technical concepts and definitions.
- Different classifications of biological databases.
- Hierarchical organization of life and levels of annotation.

Practical session #1

- Introduction to the NCBI and the *Entrez* system
- Tools for online databases search and retrieval (part I).

What is a database?



Dictionary definition

Database : A usually large collection of data organized specially for rapid search and retrieval. (as by a computer)

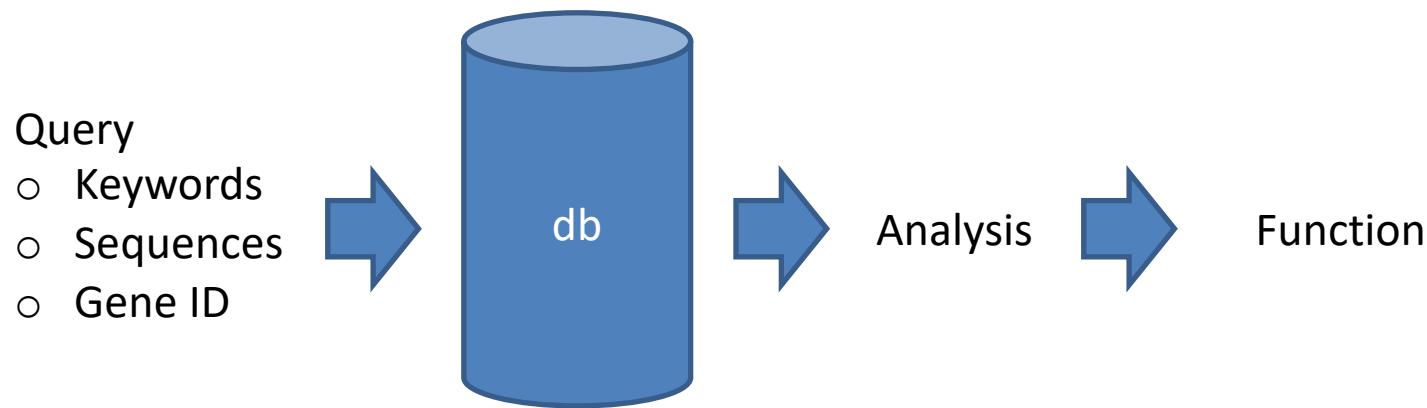
The Oxford English dictionary cites a 1962 technical report as the first to use the term "data-base."

What is a database?

A collection of related data, which are:

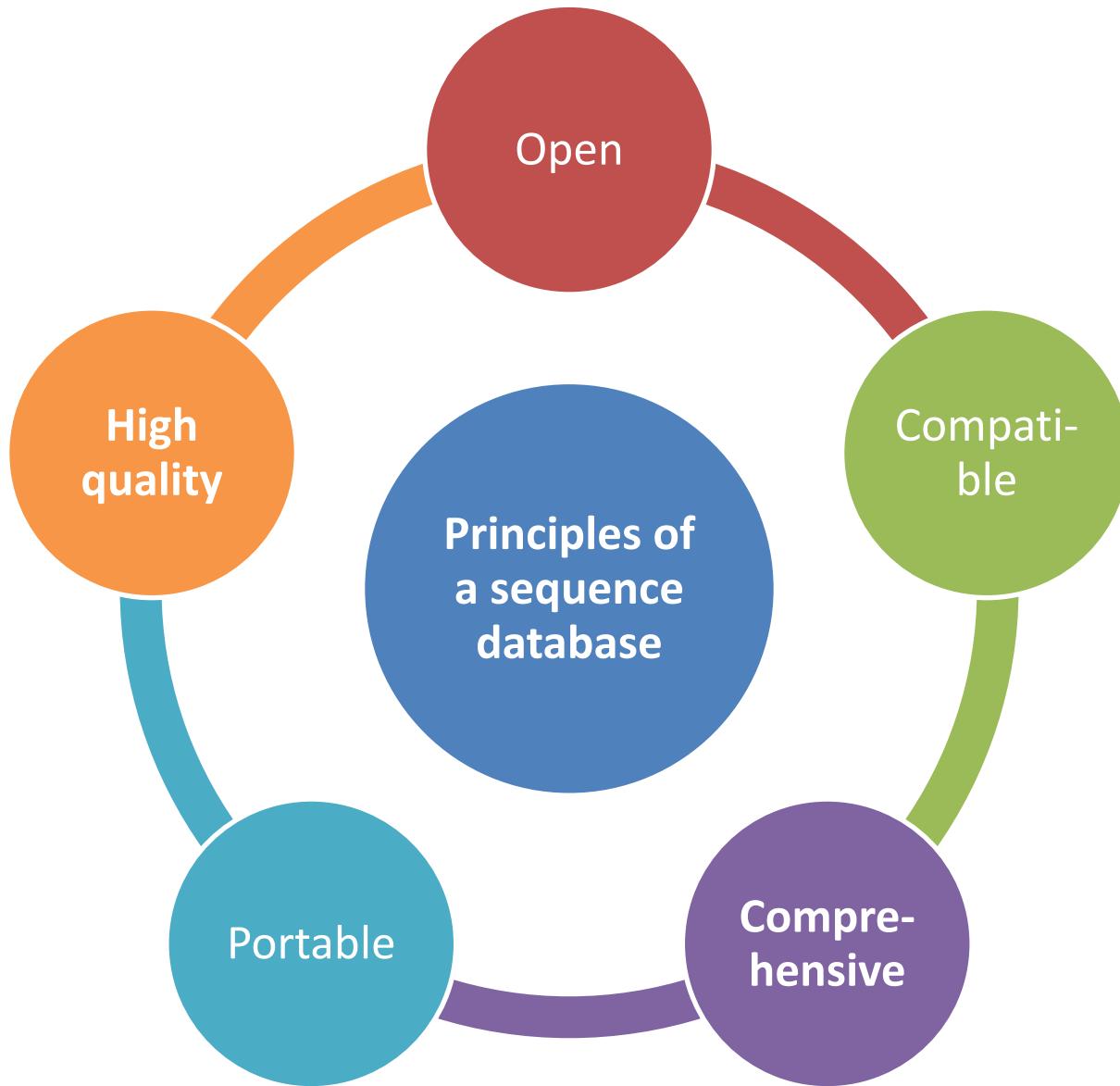
- Structured
- Searchable (index)
- Updated periodically (releases)
- Cross-referenced (hyperlinks)

You will need an appropriate database management system (DBMS)!



How Databases should be?

Taken from <http://www.ebi.ac.uk/services>



What is a database?

- A collection of related data elements
 - *Tables*
 - *columns (fields)*
 - *rows (records)*
 - *Documents*
 - *Key -> Value*
- Records retrieved using a query language
- Database technology is well established

Relational Databases

Rows (records)

- actual data
- whereas *fields* describe what data is stored, the *rows* of a table are where the actual data is stored

Columns (fields)

- attributes of tables, e.g. for *citation* table, *title*, *journal*, *volume*, *author*

How is information organized in databases?

Accession numbers and Identifiers

An **Identifier** is essentially a name of a database, table, or table column.

- As the creator of the database, you are free to identify these objects as you please.
- The identifier can change (based on the curator)

Each record (row in the table) has a **unique identifier**, alone or combined with another column is unique for that table.

The primary key (accession number or accession code).

- The primary key should not change.
- Data is indexed according this primary key
- The unique identifier serves to identify the data stored in this record across all the tables in the database (relational database).
- Usually, a string of letters and digits that uniquely identifies an entry in its database.
 - The accession number for TPIS_CHICK in Uniprot/Swissprot is P00940

How is information organized in databases?

Accession numbers and Identifiers

Some DBs have both Identifiers and accession as unique for each entry in the DB.

In these cases, the main difference is that Identifiers are Human readable and accession are just "random" codes.

Example Pfam_ID: MlaC; Pfam_AC: PF05494

In some cases Identifiers are mutable (but remain unique) while accessions are not
in UniprotKB accession P0AAP1 Use to have ADRA_ECOLI as Id (Name) but no it is
DGCC_ECOLI

The fact that they are called "unmutable" does not mean that can not be considered
obsolete and removed from the database.

Sometimes accessions can have a version number, which means that something has
changed, but whatever they represent remains

Examples:

In Pfam the current accession with version for the MlaC family is PF05494.15

In ncbi refseq nucleotide, the current accession for *Neisseria meningitidis* MC58
complete genome is NC_003112.2

Biological Databases

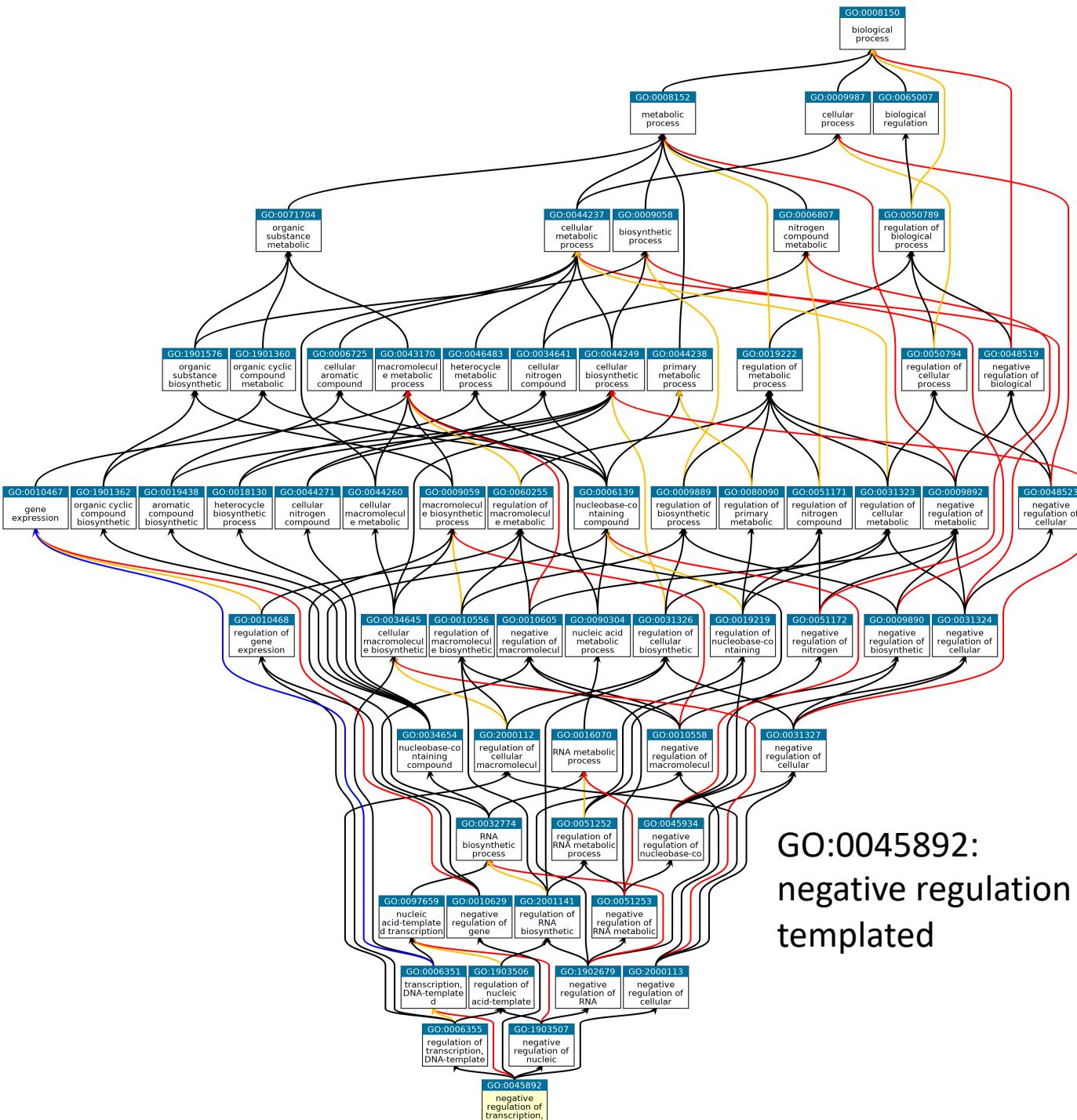
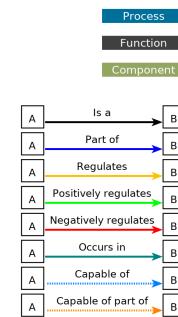
- Thousands biological databases
 - Many of the major ones covered in the annual Database Issue of Nucleic Acids Research (NAR) (2018: 1737 listings 2022:1645). It is available at <http://www.oxfordjournals.org/nar/database/c/>
- Vary in size, structure, quality, coverage, level of interest, data origin ...
- Generally accessible through the web

Limitation and challenges of biological databases

- Each of the database resources contains a different subset of biological knowledge.
- There is not standard format
 - Every database or program has its own format for storing or presenting data (eg: <http://current.geneontology.org/ontology/go-basic.obo>)
- There is not standard nomenclature
 - Every database has its own names (Controlled vocabularies)
- Data is not fully optimized
 - Some datasets have missing information without indication of it
- Data errors
 - Data is sometimes of poor quality, erroneous, misspelled
 - Error propagation resulting from computer annotation
- The integration of biological data remained an additional challenge
 - Different DBMS
 - Name of biological objects across databases (Controlled vocabularies)
 - Biological databases are continuously changing
 - Clash of concepts

Controlled vocabularies are a vital ingredient for annotating data stored in databases.

- Non-hierarchical controlled vocabularies
 - the simplest type of controlled vocabularies are non-hierarchical lists of terms
- Thesaurus as a controlled and structured vocabulary in which concepts are represented by terms
- Taxonomy as a classification scheme
 - you can find everything annotated as a sub-category of the search term
- Using ontologies
 - an ontology describes the categories of objects described in a body of data, the relationships between those objects, and the relationships between those categories (Directed Acyclic Graph)
 - *E.g.* the Gene Ontology (GO) describes the function and cellular localization of gene products across all species. Eg GO:0045892



GO:0045892:
negative regulation of transcription, DNA-templated

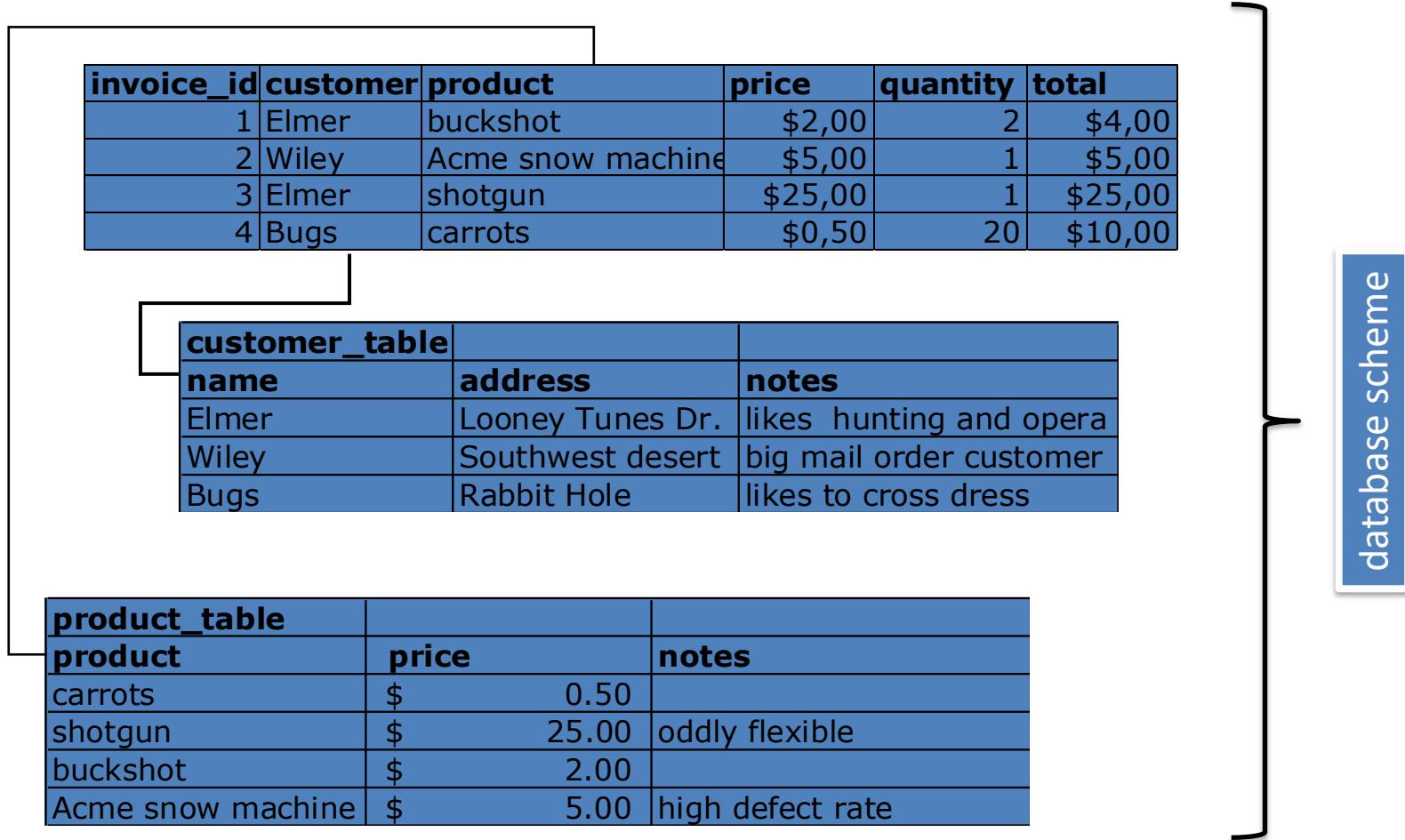
What is a flat file database?

- Sequential collection of entries, stored in a set of text files
- Flat-File databases can be represented as holding all of their data in one table only (two-dimensional table)
- Files written in plain text, standard defined format. Examples:
 - Each line is a record. Fields are separated by delimiters: tabs, commas...
 - Each file is a record. Fields expressed as key->value (eg: json db)
- Searching issues!

Accession	Source	Gene	Mol Type
AF068625.2	<i>Mus musculus</i>	dnmt3a	mRNA
HD654844.1	<i>Homo sapiens</i>	hba1	mRNA
AD836734.3	<i>Escherichia coli</i>	recA	DNA
BD823723.5	<i>Homo sapiens</i>	hpo3	DNA
TF7823562.1	VIH	p17	cDNA
AS9832656.3	<i>Homo sapiens</i>	hbb	DNA
AF6723523.1	<i>Danio rerio</i>	esf2	mRNA

What is a relational database?

- A relational database contains multiple tables and defines the relationships between them.
- Virtually all use SQL (Structured Query Language) as a language for querying and maintaining



A common way of storing biological data in a structured manner is to use a relational database

```

GeneBank Flat File Format
GeneBank Flat File Format

GeneBank Flat File Format

GeneBank Flat File Format

LOCUS AF068625 200 bp mRNA linear ROD 06-DEC-1999
DEFINITION Mus musculus DNA cytosine-5 methyltransferase 3A (Dnmt3a) mRNA,
complete cds.
ACCESSION AF068625 REGION: 1..200
VERSION AF068625.2 GI:6449467
KEYWORDS .
SOURCE Mus musculus (house mouse)
ORGANISM Mus musculus Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;
Muroidea; Muridae; Murinae; Mus.
REFERENCE1 (bases 1 to 200) , AUTHORS, TITLE, JOURNAL, etc.
REFERENCE2 (bases 1 to 200) , AUTHORS, TITLE, JOURNAL, etc.
REMARK Sequence update by submitter
COMMENT On Nov 18, 1999 this sequence version replaced gi:3327977.
FEATURES Location/Qualifiers
source 1..200 /organism="Mus musculus"/mol_type="mRNA"/db_xref="taxon:10090"
/chromosome="12"/map="4.0 cM"
gene 1..>200 /gene="Dnmt3a"
ORIGIN 1 gaattccggc ctgctccgg gcccggccac ccgcggggcc acacggcaga gccgcctgaa 61
gcccagcgct gaggctgcac tttccgagg gcttgacatc agggtctatg ttaagtctt 121 agctctgtct
tacaaagacc acggcaattc ctctctgaa gcccctcgagccccacagcg 181 ccctcgacgc cccagccgtc//
```

tab1

Accession	Source	Gene	Mol Type
AF068625.2	<i>Mus musculus</i>	dnmt3a	mRNA
HD654844.1	<i>Homo sapiens</i>	hba1	mRNA
AD836734.3	<i>Escherichia coli</i>	recA	DNA
BD823723.5	<i>Homo sapiens</i>	hpo3	DNA
TF7823562.1	VIH	p17	cDNA
AS9832656.3	<i>Homo sapiens</i>	hbb	DNA
AF6723523.1	<i>Danio rerio</i>	esrf2	mRNA

tab2

Species	TaxID	Synonym
<i>Homo sapiens</i>	9606	Human
<i>Mus musculus</i>	10090	Mouse
<i>Danio rerio</i>	7955	Zebra fish
<i>Escherichichia coli</i>	562	E. coli

Essential aspects of primary and secondary databases.

	Primary database	Secondary database
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers (database staff organize but don't add additional information)	Results of analysis, literature research and interpretation, often of data in primary databases
	Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.	Continuously updated <u>Biocuration</u>

EMBL-EBI Train online
Bioinformatics for the terrified

<https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/>

Definition and aims of biocuration

Biocuration involves the interpretation and integration of information relevant to biology into a database or resource that enables integration of the scientific literature as well as large data sets.

Primary goals of biocuration.

- Accurate and comprehensive representation of biological knowledge
- Easy access to this data for working scientists and a basis for computational analysis

How to access the data ?

Databases Search and Retrieval

A request of data from a Database is called as Query

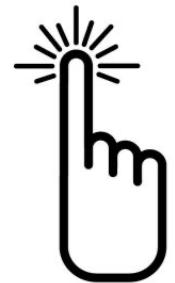
Queries can be of three forms:

1. Choose from a list of parameters
2. Query by example (QBE)
 - QBE build wizard allows which data to display
3. Query language
 - SQL (structured query language)

How to access the data in public databases ?

Human Web interface (web based, small scale)

- Free text search
- Common mode of search are keywords with modifiers or identifiers
- Cross-references link the information of different databases
- You do not see the underlying database structure
- Output defined by host/provider



“click your way”

Web services and Programmatic data access

- Application Programmers Interface (API)
- To approach database programmatically



Programming Utilities Web Service

Download the data: File Transfer Protocol (FTP), rsync, http

- Flat files (script based, bulk data download)

Database searching tips to choose from a list of parameters

- Using **keywords** and enclose phrases in **double quotes**
- **Boolean** searches

Boolean logic consists of three logical operators:

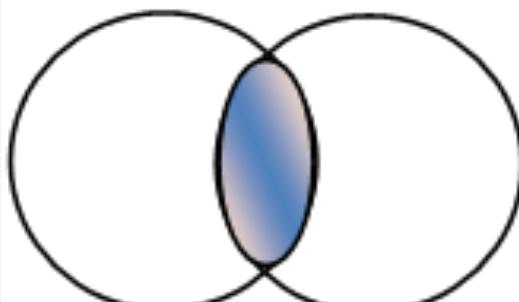
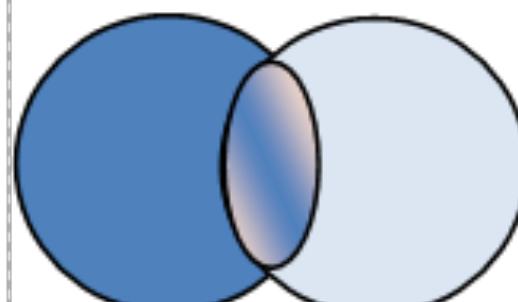
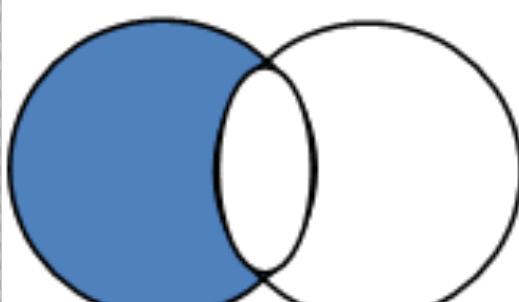
OR

AND

NOT

- Wildcards or Query Truncation
- **Advance searches** by using search tags and **fielded searching**

Searching Strategies: Boolean operators

cats AND dogs	cats OR dogs	cats NOT dogs
		
retrieves items that contain both terms	retrieves items that contain either term	retrieves items that contain only one term
Restrict	Expand	Filter

Searching with wildcards or query truncation.

- Truncation: Wildcards are useful if, for example, you wish to search for a group of words (e.g., all words starting with “cell” and ending with “ase”) or if it is unclear how a word is spelt in a databank.
 - Cell* (NCBI and EMBL)
 - *ase (EMBL)
 - *moglobin (EMBL)
- “?” Matches one character of any value.
 - nif? (EMBL)
This expression finds the gene names nifa, nifb, nifc, nifd, nife. But do not words like Nifedipine

Note: Placing a wildcard at the start of a word or string may increase the response time because all words in the index have to be checked against your string.

For example cat* in PubMed, will give incomplete results!

Fielded searching using any of the indexed fields (advanced searches)

Entering the phrase with a field descriptions:

robotic surgery [title]

Miller MJ [author]

“protein domain” [TI]

human [Organism]

insulin [Protein Name]

Combining fielded searching with booleans

enzymes [TI] NOT Gonzales P [AU]

human [Organism] AND insulin [Protein Name]

Search for Field Descriptions are different in each Database

NCBI

NC_0000*[Accession]
Human[Organism]
horse[taxonomy]
neoplasms[MeSHTerms]
prolactin[Protein Name]
APOE[gene]
srcdb_refseq[Properties]
2010/06[Publication Date]
110:500[Sequence Length]
gene_symbol[sym]
1.1.1.53[ecno]
gbdiv_est[PROP]
:
etc

UNIPROT

accession:p62988
organism:human
taxonomy:40674
keyword:neoplasms
name:"prion protein"
gene:HSPC233
database:(type:pfam)
created:[20121001 TO *]
length:[100 TO 500]
go:0015629
ec:3.2.1.23
reviewed:yes
:
etc

http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

<https://www.ncbi.nlm.nih.gov/books/NBK49540/>

<https://www.uniprot.org/help/query-fields>

The NCBI is a comprehensive website for biologists (database of databases (of databases))

<http://www.ncbi.nlm.nih.gov/gquery/>

- The National Center for Biotechnology Information (NCBI)
- Created in 1988 as a part of the National Library of Medicine at NIH
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information
- Over 30 databases (primary, secondary, specialized, meta-databases, etc.)



The NCBI home page

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

NCBI Resources How To masterbioinf My NCBI

NCBI National Center for Biotechnology Information All Databases Search

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

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

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

NCBI Announcements

- Genome Workbench 2.10 now available
- Genome Workbench 2.10 includes a reworked BLAST tool and new functionalities in Tree View. For more details, see the announcement.
- Sequence Viewer 3.11 now available
- Sequence Viewer 3.11, now available, contains a number of new features and improvements and bug fixes. For more details, see the announcement.

NCBI hosts over 30 databases

NCBI Resources How To



National Center for
Biotechnology Information

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

All Databases

- MeSH
- NCBI Web Site
- NLM Catalog
- Nucleotide**
- OMIM
- PMC
- PopSet
- Probe
- Protein
- Protein Clusters
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed Health
- SNP
- Sparcle
- SRA
- Structure
- Taxonomy

to NCBI

Center for Biotechnology Information advances science and providing access to biomedical and genomic information.

[NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Download

Transfer NCBI data
to your computer



Learn

Find help
documents, attend
a class or watch a
tutorial



How to access the NCBI data ?

***Entrez*: An Integrated Database Search and Retrieval System**

Human Web interface (web based, small scale)

- Free text search
- List of identifiers (Batch *Entrez*)

Web service (Programmatic data access)

- *Entrez Utilities Web Service (NCBI): The E-utilities*
 - *Entrez Direct: E-utilities on the Unix Command Line*

Searching sequence databases using a sequence query

- BLAST

File Transfer Protocol (FTP)

- Flat files (script based, bulk data download)

***Entrez*: An Integrated Database Search and Retrieval System**

- Access all NCBI resources (Database Integration)
- *Entrez* Databases
 - All Molecular Database entries are organized by organism (Taxonomy Database).
 - Each record is assigned a UID “unique integer identifier” for internal tracking
 - Each record is indexed by data fields: [author], [title], [organism], and many others
 - Each record is given a Document Summary (DocSum).
 - Each record is manually or computationally assigned links to biologically related UIDs in and across databases.



The screenshot shows the NCBI Batch Entrez search interface. At the top left is the NCBI logo. To its right is the title "Batch Entrez". Below the title is a horizontal menu bar with links: All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, and Books. Under the "Database" dropdown, "Nucleotide" is selected. Below the menu is a search bar with a file input field labeled "File:" and a "Choose File" button. The input field shows "no file selected". To the right of the input field is a "Retrieve" button.

Batch Entrez

Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

Instructions

1. Start with a local file containing a list of accession numbers or identifiers
2. Select the database corresponding to the type of accession numbers or identifiers in your input file
3. Use the **Browse** or **Choose File...** button to select the input file
4. Press the **Retrieve** button to see a list of document summaries
5. Select a format in which to display the data for viewing, and/or saving
6. Select 'Send to file' to save the file.

Tips

- To download entire genome records, check the NCBI FTP site, instead of using Batch Entrez.
- Some lists of record identifiers can be tens of thousands of lines long, so Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems.
- When loading large numbers of genome records, put several thousand record identifiers per file, one per line, left-adjusted.
- Please note that Batch Entrez will check for duplicate identifiers when reporting results from a list that you have imported.
- When retrieving a list of Nucleotide accessions, you must select the specific component database from which the accessions or GIs were saved. For Nucleotide, choose either the CoreNucleotide, the EST or the GSS selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.
- In all cases, be certain to select the database that corresponds to the identifiers you are uploading. For example, if you have saved a list of protein accession numbers, be sure to select the Protein database.

Access Entrez through programs or scripts

Entrez Utilities Web Service (NCBI) **The E-utilities** <http://www.ncbi.nlm.nih.gov/books/NBK25500/>

- Entrez Programming Utilities are tools that provide access to Entrez data outside of the regular web query interface. You can access them by constructing the right URL and retrieving results. This is the base:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>

And then you add the desired tool with the appropriate parameters.

Some tools:

- **ESearch: Searches and retrieves primary IDs and retains results in the user's environment.**
- **EFetch: Retrieves records from one or more primary IDs or from the user's environment.**
- Also: **EGQuery, EInfo, ELink, ESpell, Esummary**

Example: Get the PubMed IDs (PMIDs) for articles about breast cancer published in Science in 2008:

[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=science\[journal\] AND breast cancer AND 2008\[pdat\]&retmode=json](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=science[journal] AND breast cancer AND 2008[pdat]&retmode=json)

- There is a set of programs implementing these functions. They allow to access the same services from the UNIX-like Command Line interface : Entrez Direct.
(<https://www.ncbi.nlm.nih.gov/books/NBK179288/>)

Example: Download all protein sequences in fasta format from *Stenotrophomonas maltophilia* with the word “lactamase” in their title:

```
~]$ esearch -db protein -query "txid40324[prorgn] AND lactamase[TI]"|efetch -format fasta
```

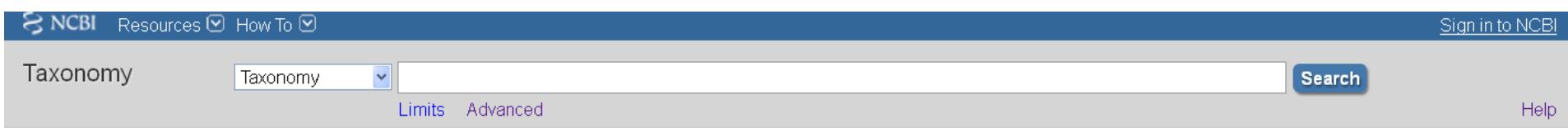
BLAST (Basic Local Alignment Search Tool)

Altschul, *et al.* 1990, *J Mol Biol*, 215:403-10*

Altschul, *et al.* 1997, *Nucleic Acids Res*, 25(17): 3389–3402
(* one of the most highly cited paper)

- BLAST is an algorithm to find regions of similarity between biological sequences both proteins or nucleic acids
- BLAST compares a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.
- BLAST Home page: <http://www.ncbi.nlm.nih.gov/BLAST>
- BLAST is one of the most widely used bioinformatics programs

NCBI: Taxonomy DataBase



NCBI Resources How To Sign in to NCBI

Taxonomy Taxonomy Search

Limits Advanced Help



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

[Quick Start Guide](#)

[FAQ](#)

[Handbook](#)

[Taxonomy FTP](#)

Taxonomy Tools

[Browser](#)

[Common Tree](#)

[Statistics](#)

[Name/ID Status](#)

[Genetic Codes](#)

[Linking to Taxonomy](#)

[Extinct Organisms](#)

Other Resources

[GenBank](#)

[LinkOut](#)

[E-Utilities](#)

[Batch Entrez](#)

[INSDC](#)

530 777 (formal) species!

Each taxa with an ID, the **TaxID**

<http://www.ncbi.nlm.nih.gov/taxonomy/>

NCBI:Molecular Sequence Databases

Sequence Databases (Primary)

Nucleotide (GenBank)
PopSet
SRA, GSS
Protein

Marker Databases

Single Nucleotide Polymorphisms (SNP's, dbSNP)
Sequence Tagged Sites (STS's, dbSTS)
Expressed Sequence Tags (EST's, dbEST)

NCBI Resources ▾ How To ▾ Sign In

Nucleotide Nucleotide Advanced



Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

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

NCBI: Derivative Databases

Nucleotide derived

Example: RefSeq, GENE

Protein-derived

Example: CDD

Structure-derived

Example: Structure

Human curated, compilation and correction of data

Example: RefSeq

Computationally Derived

Example: UniGene

Combinations

Example: NCBI Genome Assembly



Resources ▾ How To ▾

[Sign in to NCBI](#)

RefSeq

RefSeq

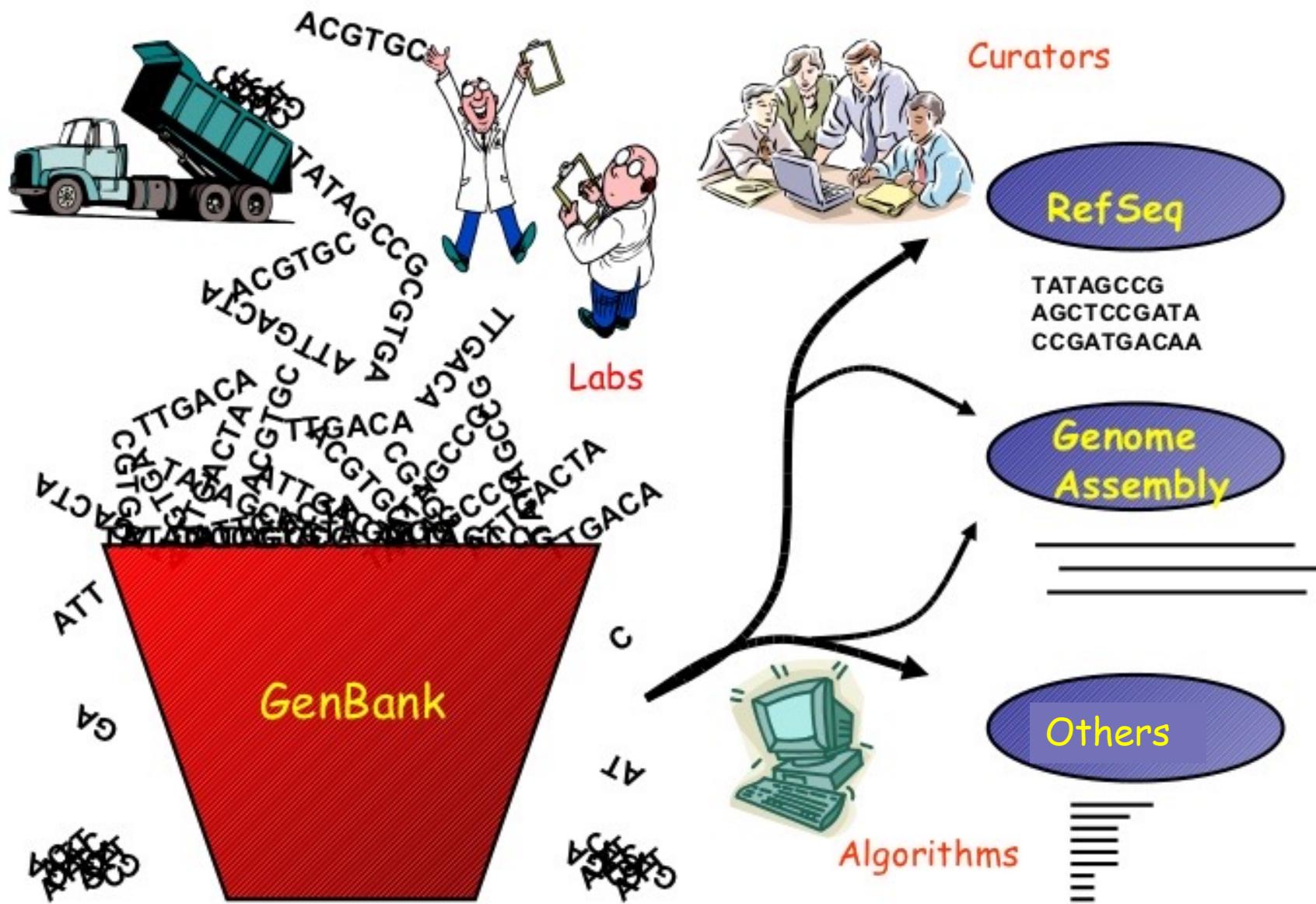
[Search](#)

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

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

NCBI: Derivative Databases



The EMBL-EBI website has been redesigned. Please [send us feedback](#) about this page.

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Example searches: [blast](#) [keratin](#) [bfl1](#) | [About EBI Search](#)

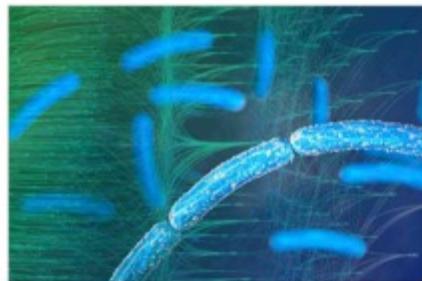
[Find data resources](#) →

[Submit data](#) →

[Explore our research](#) →

[Train with us](#) →

Latest news →



Shining a light on how bacteria interact

01 Aug 2022



AlphaFold predicts structure of almost every catalogued protein known to science



Uncharted territories in the human genome

13 Jul 2022



Ensembl 107 has been released

12 Jul 2022

The EMBL-EBI's search engine

The EMBL-EBI nucleotide repository:ENA

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

The screenshot shows the ENA homepage with a dark header containing the EMBL-EBI logo and links for Services, Research, Training, About us, and a search icon. Below the header is the ENA logo and the text "European Nucleotide Archive". The main menu includes Home, Submit, Search, Rulespace, About, and Support. To the right are two search boxes: one for text search terms (with examples like histone, BN000065) and another for accession numbers (with examples like Taxon:9606, BN000065, PRJEB402). A message encourages subscribing to the ENA-announce mailing list. Another message provides information about SARS-CoV-2 data submissions.

We recommend that you subscribe to the [ENA-announce mailing list](#) for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#).

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.



A screenshot of a Twitter interface showing a tweet from the account @ENASequencing. The tweet is from Sundar Pichai (@sundarpichai) and discusses the expansion of the AlphaFold Protein Structure Database. It includes a link to a DeepMind tweet.



The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure
The ENA is an ELIXIR Core Data Resource. [Learn more](#) >

The mission of **UniProt** is to provide the scientific community with a comprehensive, high quality and freely accessible resource of protein sequence and functional information.

Find your protein

UniProtKB ▾

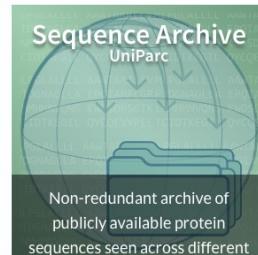
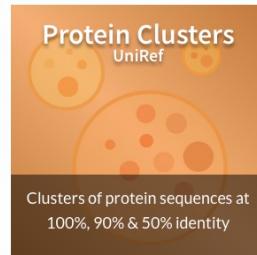
Examples: Insulin, APP, Human, P05067, organism_id:9606

Advanced | List Search

Feedback

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt™](#)

ⓘ Accessing UniProt programmatically? Have a look at the [new API documentation](#). If you still need it, the legacy version of the website is available until the 2022_03 release.



Taxonomy
Literature Citations

Subcellular locations
Cross-referenced databases

UniRule automatic annotation
ARBA automatic annotation

Keywords

Ever had to deal with tiny polystyrene balls clinging to your carpet? Or admired the neat narrow grooves stamped onto a vinyl record? Perhaps you have just acquired a sleek viscose shirt. Polystyrene, PVC, vinyl...

Metrics, metrics, metrics... UniProt is brought to you by a large team of dedicated scientists who have worked for over 2...

The UniProt Metal Bin... We would like to invite the machine learning community to help UniProt by creating...

UniProt Data Explorer



UniProt Data Explorer

Metrics, metrics, metrics... UniProt is brought to you by a large team of dedicated scientists who have worked for over 2...

The UniProt Metal Bin... We would like to invite the machine learning community to help UniProt by creating...

Latest News

UniProt release 2022_03
Not just for proteins: new targets for ADP-ribosylation | Annotation of biologically relevant...

UniProt release 2022_02
Prenylation for antiviral activity | Cross-references to AlphaFoldDB | Version numbers f...

<http://www.uniprot.org/>

REST Web services access:

<http://www.uniprot.org/help/query-fields>

https://www.uniprot.org/help/api_queries

https://www.uniprot.org/help/return_fields

https://www.uniprot.org/help/api_retrieve_entries

If you want more:

<https://www.uniprot.org/help/api>

← → ⌂ https://www.genome.jp/kegg/

Getting Started ARE - UAB Curs: Bioinformàtic... Course: Public Data... 2020 UPF SIGMA ACADEMIC MSc in Bioinformati...

KEGG Databases Mapper Auto annotation Kanehisa Lab



KEGG Search Help » Japanese

KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (August 1, 2021) for new and updated features.

New article KEGG mapping tools for uncovering hidden features in biological data

Main entry point to the KEGG web service

KEGG2 KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENES	Genes and proteins [SeqData]
KEGG GENOME	Genomes [KEGG Virus Taxonomy]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

Organism-specific entry points

KEGG Organisms Enter org code(s) Go hsa hsa eco

Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)
Antimicrobials

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

REST API:

<https://www.kegg.jp/kegg/rest/keggapi.html>

<https://www.rcsb.org>

https://www.rcsb.org/

Web services

<https://www.rcsb.org/docs/programmatic-access/web-services-overview>

Doc: <https://data.rcsb.org/redoc/index.html>

Data API tutorial: <https://data.rcsb.org/index.html>

Search API tutorial: <https://search.rcsb.org/index.html>

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB

181969 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search terms or PDB ID(s). Advanced Search | Browse Annotations Help

PDB-101 Worldwide Protein Data Bank EMDDataResource NUCLEIC ACID DATABASE Worldwide Protein Data Bank Foundation

Celebrating 50 YEARS OF Protein Data Bank

Developers: Join the RCSB PDB Team Explore Open Positions

Welcome Deposit Search Visualize Analyze Download Learn

A Structural View of Biology

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

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data. The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

COVID-19 CORONAVIRUS Resources

Celebrating PROTEIN DATA BANK 50 Years

September Molecule of the Month

DNA-Sequencing Nanopores

Screenshot of the SIFTS homepage (<https://www.ebi.ac.uk/pdbe/docs/sifts/>)

The page shows the EMBL-EBI logo and a navigation bar with links to Services, Research, Training, and About us.

SIFTS

SIFTS home Quick access Overview Methods Statistics XML schema Acknowledgments

Structure integration with function, taxonomy and sequence

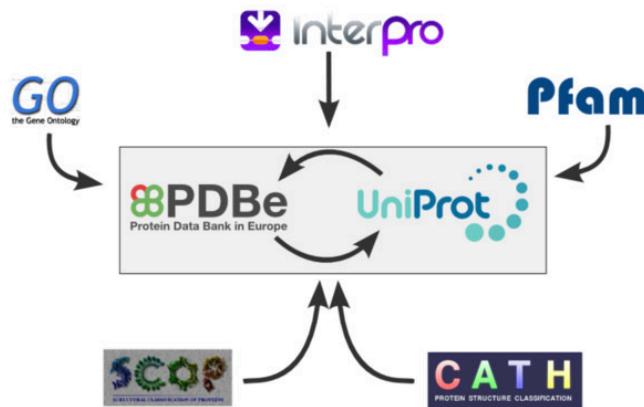
Structure Integration with Function, Taxonomy and Sequence (SIFTS) is a project in the PDBe-KB resource for residue-level mapping between UniProt and PDB entries. SIFTS also provides annotation from the IntEnz, GO, InterPro, Pfam, CATH, SCOP, PubMed, Ensembl and Homologene resources. The information is updated and released every week concurrently with the release of new PDB entries and is widely used by resources such as RCSB PDB, PDBj, PDBeSum, Pfam, SCOP and InterPro.



If you use SIFTS, please cite:

Dana, Gutmanas et al., Nucleic Acids Research 47, D482 (2019)

Velankar et al., Nucleic Acids Research 41, D483 (2013)



<https://www.ebi.ac.uk/pdbe/docs/sifts/>

<https://www.ebi.ac.uk/pdbe/docs/sifts/overview.html>

REST API: <http://www.ebi.ac.uk/pdbe/api/doc/sifts.html>

