Nucleotide and protein databases

Adapted from the courses of the Bonsai team,

CRIStAL UMR 9189

Sylvain.legrand@univ-lille.fr

Introduction

Definitions

- A field of research that **analyses and interprets biological data**, using **computational methods**, to create new knowledge in biology (Quninkal and Rechenmann, 2004)

- In English, there are two terms:
  - **Bionformatics**: applies algorithms, statistical models with the aim of interpreting, classifying and understanding biological data
  - **Computational Biology**: developing mathematical models and associated tools to solve biological problems

- In French: “Bioanalyse” → Bionformatics; “Recherche en bioinformatique” → Computational Biology

Definitions

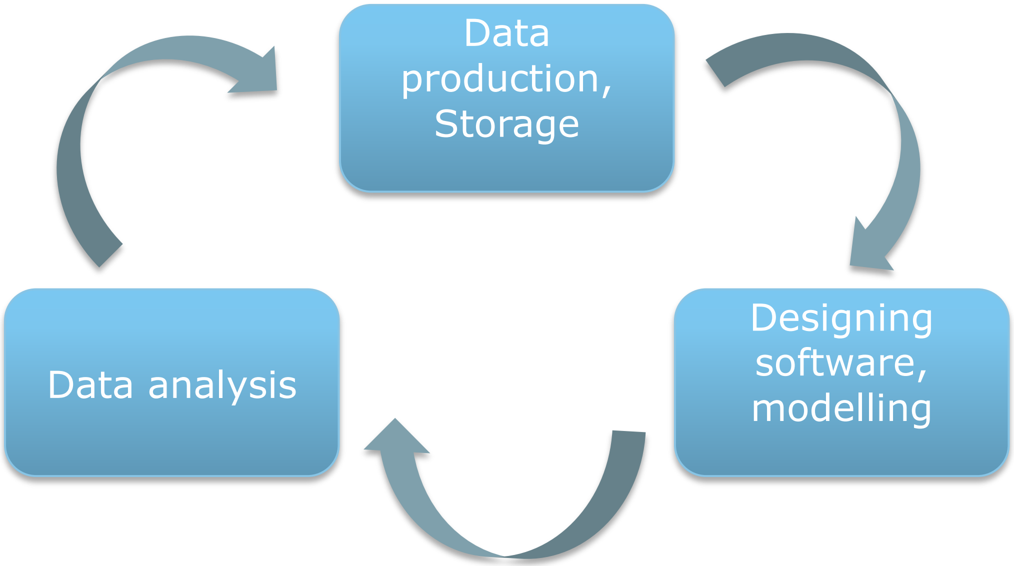
- **A simple definition**: the *in silico* approach to biology

Biology

Bioinformatics

Computer science

- **Three main activities**



Data production, Storage

Designing software, modelling

Data analysis

Tips

- **Beware** of software **results**

  - The quality of the results is sometimes diminished in favour of the speed

  - Some problems admit an infinite set of possibilities → it is not always the best solution that is found

  - Some software provide only predictions

- **Beware** of **databases**

  - Data are not always reliable

  - Data updates are not always recent

Some fields of application

- **Bioinformatics of biological sequences**

  DNA, proteins, sequence alignment, gene identification...

- **Bioinformatics for metabolites**

  Identification, annotation...

- **Structural bioinformatics**

  Analysis of the folding of biological macromolecules

- **Bioinformatics of networks**

  Interactions between genes, proteins, metabolites...

- **Bioinformatics for population genetics**

  Ex: modelling the evolution of populations in specific environments...

Database, definition

- **Set of data** relating to a domain, **organized by computer** processing, **accessible** on-line

- Generally, data are stored as **formatted text files** (with a special layout)

- Need to develop **specific software to query** the data contained in these databases

- **3 main centres** to collect and to provide freely available sequences : EBI, NCBI, DDBJ

In Europe: EBI

- **European Bioinformatics Institute**
  http://www.ebi.ac.uk/

- Non-profit academic organization founded in **1992**

- Bioinformatics research and services centre that manages biological databases (DNA-RNA, proteins, 3D structures)

- Places in the public domain and makes accessible free of charge the information resulting from research in molecular biology and genomics in order to promote the scientific advancement

EMBL-EBI

In USA: NCBI

- **National Center for Biotechnology Information** http://www.ncbi.nlm.nih.gov/
- Founded in **1988**
- Creation of public database and research in bioinformatics
- Develops, distributes, supports, and coordinates access to a variety of databases and software for the scientific and medical communities

NCBI

In JAPAN: DDBJ

- **DNA Data Bank of Japan** https://www.ddbj.nig.ac.jp/index-e.html
- Founded in **1986**
- DDBJ Center collects nucleotide sequence data and provides freely available sequence data and supercomputer system, to support research activities in life science.

DDBJ
DNA Data Bank of Japan

Examples of formats

- **Database format**
  - DNA/RNA sequences: EMBL, GenBank, DDBJ
  - protein sequences: SwissProt, TrEMBL…

- Sequence formats used by almost all bioinformatics tools
  - **Fasta**
  - Raw sequence

- **Format conversion**
  - When viewing or downloading data
  - Example: Seqret (EBI), a tool to convert a format in an other one
  https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/
  (web or offline)

Fasta

- Used by sequence analysis software
- **One line of ID/comments** preceded by “**>**”
- The **raw sequence** (no space, no number)

```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds
ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccat
agacttgcgctcggtcaagagccgcagcgaggccggggagccgcccagct
ccctccaggtgaagcccgagacaccggcgtcggcggcggtggcggtggcg
Gcggcagcggcacccaccacgacggcggagaagcct
>hPc2 gene
ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata
taattatcaccgacgtcaccgcgaactgcctcaccgttactttcaaggag
tacgtgacggtg
```

Nucleotide databases

Nucleotide databases

- **Origin of the data**: **sequencing** of DNA or RNA molecules

- **Stored data**: 1 sequence + its annotations = 1 record (entry)

  - Fragment of genome: one or more genes, a gene fragment intergenic sequence, ...

  - Complete genome

  - mRNA, tRNA, rRNA, ... (fragments or whole molecule)

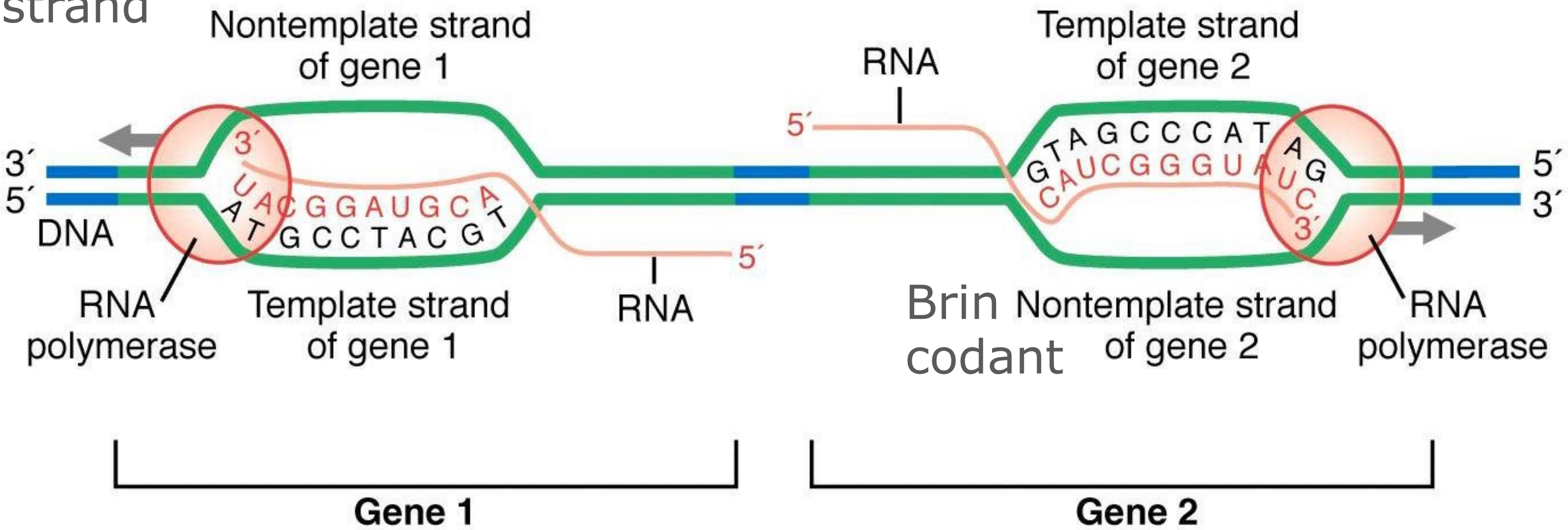
- All of the sequences (DNA or RNA) are **written with "T"**

- The strand given in the database is called the **"+" or direct strand**.

  Be careful, it is not necessarily the coding strand !

Nucleotide databases

Coding strand



Brin codant

\>Sequence
NNNNATGCCTACGTNNNNNNNNNCATCGGTATCNNNNNNNNN

Griffiths et al 2002

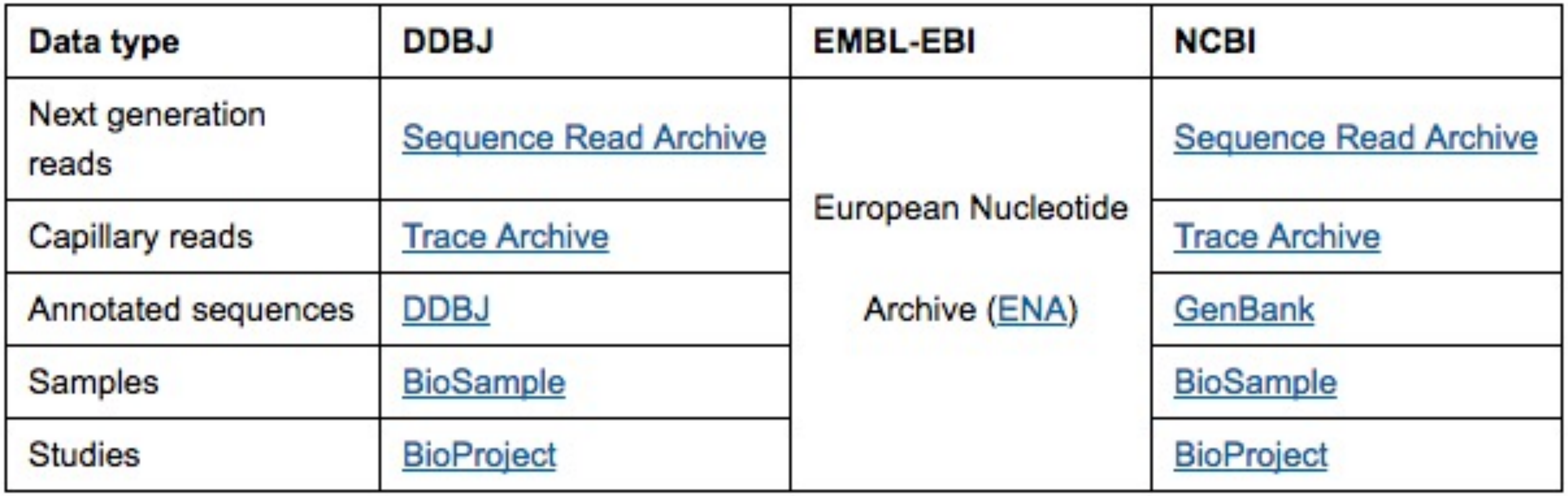
Nucleotide databases, collaboration

INSDC International Nucleotide Sequence Database Collaboration

=

DDBJ DNA Data Bank of Japan + ENA European Nucleotide Archive + NCBI

Daily exchange of data between the 3 banks



Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Nucleotide databases, updates

- **New versions** available several times a year
  - Date and version number (release)
  - Data frozen at a fixed date (the sequences collected so far)

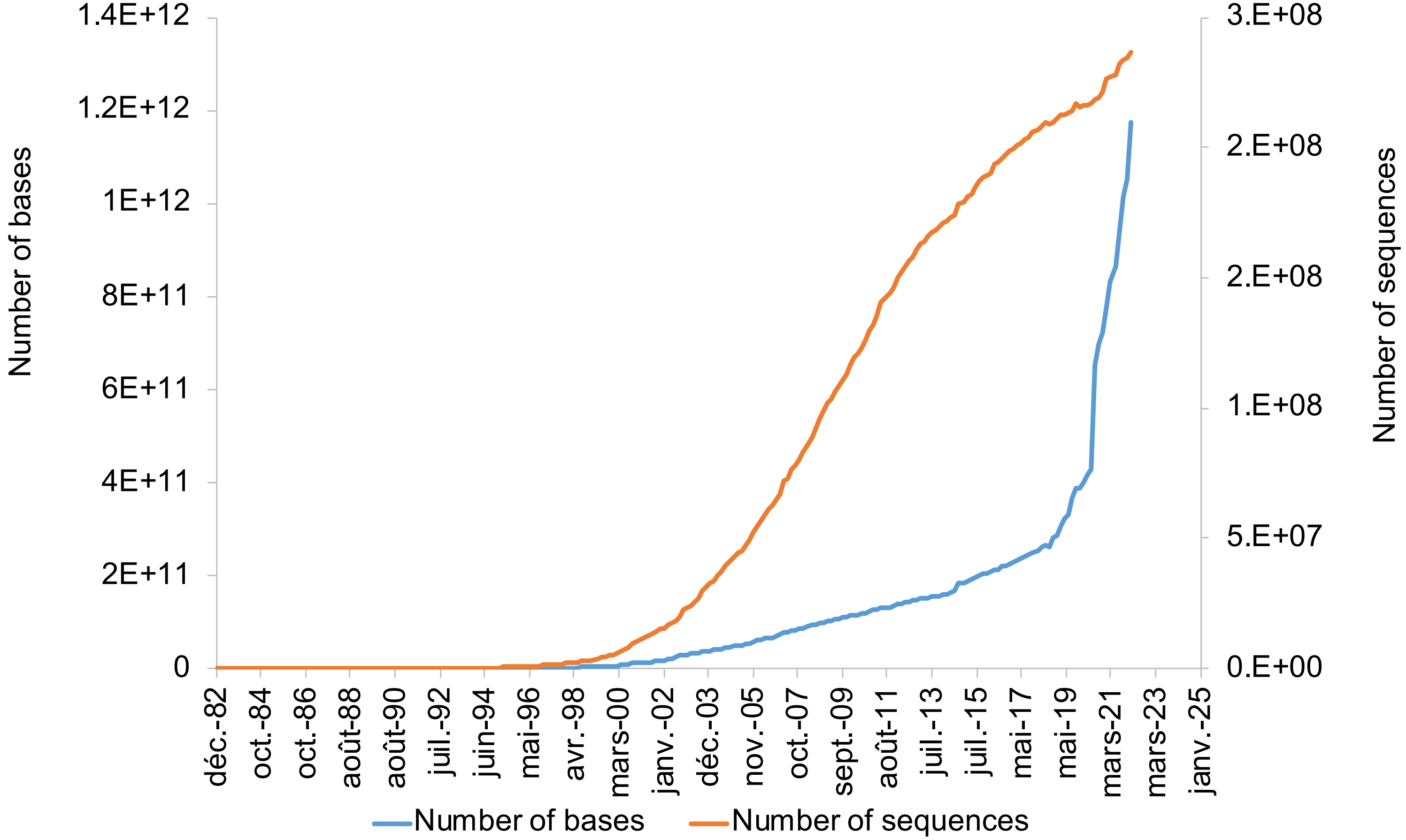
- **Updates**
  - Daily update of data
  - All new sequences since the last release
  - mRNA, tRNA, rRNA, ... (fragments or integers)

- **Ease of data processing**

- No need to download the entire bank with each update

Nucleotide databases, increase in number of sequences

GenBank size in march 2022 (genbank/statistics/)



Université de Lille

Format of a record

- **3 parts**:

- Main description

- Features: description of biological objects on the sequence

- Sequence

```
1   cttctctccc cctccaacga tctccacctc caattttcaa accctaattc tcctgttttt
61  tttatattac ttacactcct tttttatttc atctctattt cttcacatcc tccatctttt
121 gtatcatttg agttgactcg gaagttggaa ttttggattt tgatttggct tagcttgtgt
181 tgtttgacgg atggtttcta gggagtaatc cgtatagatt atgggcagta gtgggatgga
```

- **Each line starts with a key**:
  - Two letters for EMBL
  - A word of max. 12 letters for GenBank and DDBJ

- **End of record**: “//”

EMBL format

- **ID**: first line, summary

Accession	Version	Topology	Molecule	Classe	Taxonomy	Length
M71283	SV 1	linear	genomic DNA	STD	PRO	1322 BP

- **AC**: accession number (unique number)
- **DT**: dates of first release and last version
- **DE**: record description
- **KW**: keywords
- **OS**, **OC**: organism and its taxonomy
- **RN**, **RC**, **RX**, **RP**, **RA**, **RT**, **RL**: bibliographical references

GenBank and DDBJ format

- **LOCUS**: first line, summary

Locus name	Length	Molecule	Topology	Taxonomy	Date
BACCOMQP	1322 bp	DNA	linear	BCT	26-APR-1993

- **DEFINITION**=DE
- **ACCESION**=AC
- **VERSION**~DT
- **KEYWORDS**=KW
- **SOURCE**, **ORGANISM**=OS, OC
- **REFERENCE**, **AUTHORS**, **TITLE**, **JOURNAL**,…=R…
- http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Features

- **Common format** for all banks

- **Key**: a single word indicating a functional group
  - Controlled and hierarchical vocabulary

  Ex: gene: whole gene sequence, including introns
  CDS: coding sequence (without introns, between start and stop codons)

- **Location**: position of the object in the sequence

- **Qualifiers**: precise description of the functional group
  - Format: /qualifier=’open comments’

  Ex: /gene=‘comQ’ → name of the gene
  /product=‘comQ’ → name of the protein
  /protein_id=‘BAB1349.1’ → protein accession number
  /note=‘competence protein Q;…’ → information related to the function

Location

- **467**: the annotation concerns a single base

- **109..1105**: between positions 109 and 1105 (included)

- **<1..21** or **1275..>1322**: truncated keys
  - starts before beginning of the sequence
  - Ends after end of the sequence (length 1322)

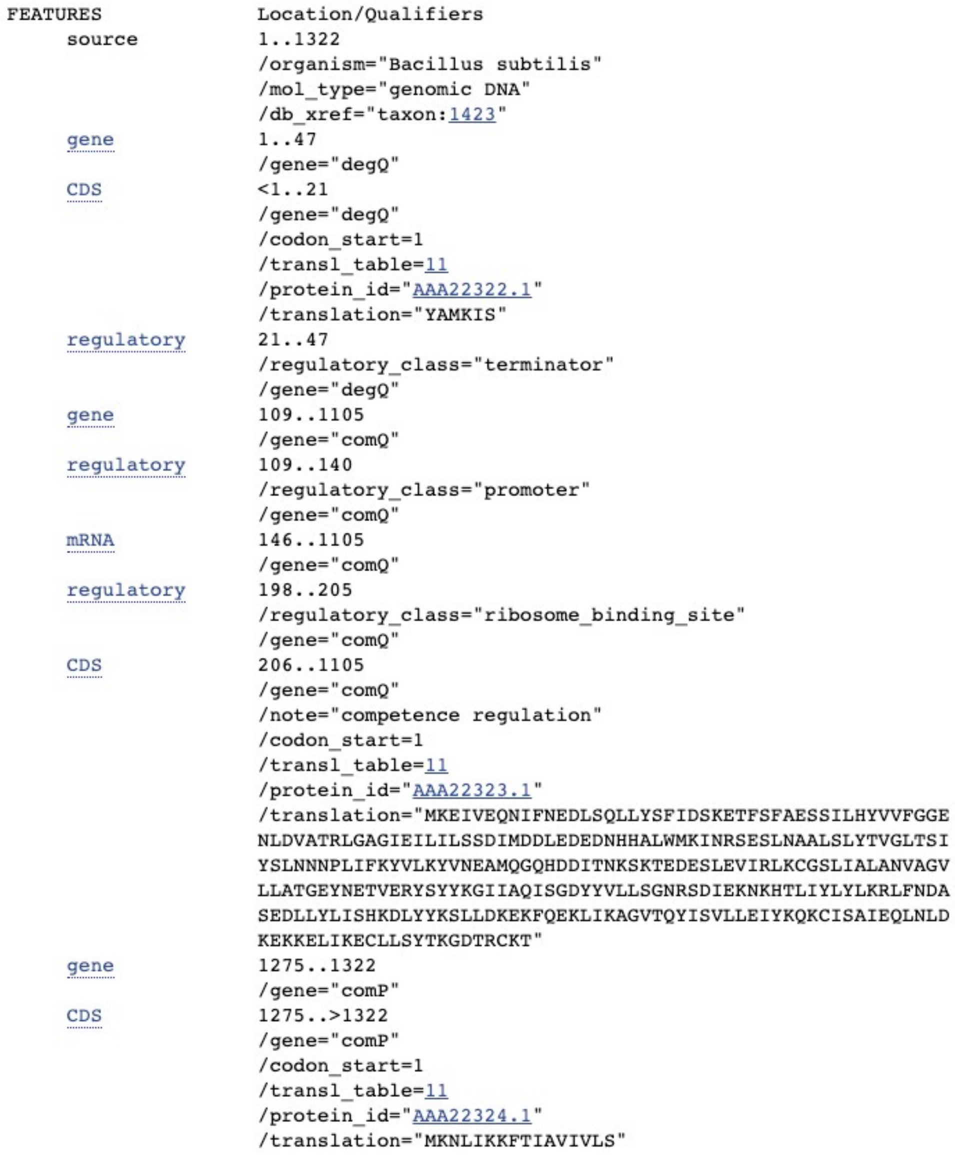
- **<234..888**: unknown start, but before 234

- **234..>888**: unknown end, but after 888

- **complement(340..565):** reverse complement of the sequence (minus strand)

- **join(12..78,134..202):** assembled fragments (concatenated), illimited number of fragments

Features, example



```
FEATURES             Location/Qualifiers
     source          1..1322
                     /organism="Bacillus subtilis"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:1423"
     gene            1..47
                     /gene="degQ"
     CDS             <1..21
                     /gene="degQ"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA22322.1"
                     /translation="YAMKIS"
     regulatory      21..47
                     /regulatory_class="terminator"
                     /gene="degQ"
     gene            109..1105
                     /gene="comQ"
     regulatory      109..140
                     /regulatory_class="promoter"
                     /gene="comQ"
     mRNA            146..1105
                     /gene="comQ"
     regulatory      198..205
                     /regulatory_class="ribosome_binding_site"
                     /gene="comQ"
     CDS             206..1105
                     /gene="comQ"
                     /note="competence regulation"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA22323.1"
                     /translation="MKEIVEQNIFNEDLSQLLYSFIDSKETFSFAESSILHYVVFGGE
                     NLDVATRLGAGIEILILSSDIMDDLEDEDNHHALWMKINRSESLNAALSLYTVGLTSI
                     YSLNNNPLIFKYVLKYVNEAMQGQHDDITNKSKTEDESLEVIRLKCGSLIALANVAGV
                     LLATGEYNETVERYSYYKGIIAQISGDYYVLLSGNRSDIEKNKHTLIYLYLKRLFNDA
                     SEDLLYLISHKDLYYKSLLDKEKFQEKLIKAGVTQYISVLLEIYKQKCISAIEQLNLD
                     KEKKELIKECLLSYTKGDTRCKT"
     gene            1275..1322
                     /gene="comP"
     CDS             1275..>1322
                     /gene="comP"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA22324.1"
                     /translation="MKNLIKKFTIAVIVLS"
```

Data updates and limits

- **Possible evolution of inputs**
  - Changes in the sequence, in the annotations
  - Adding a sequence, an annotation, a publication
- Records are **updated by their authors only**.
- **High redundancy**: same sequence fragment present in several records
- **Poorly standardized annotations:** difficulty in searching for a particular piece of information
- (Often) **imprecise annotations:** few descriptions of genes and their products
- **Errors** in annotations

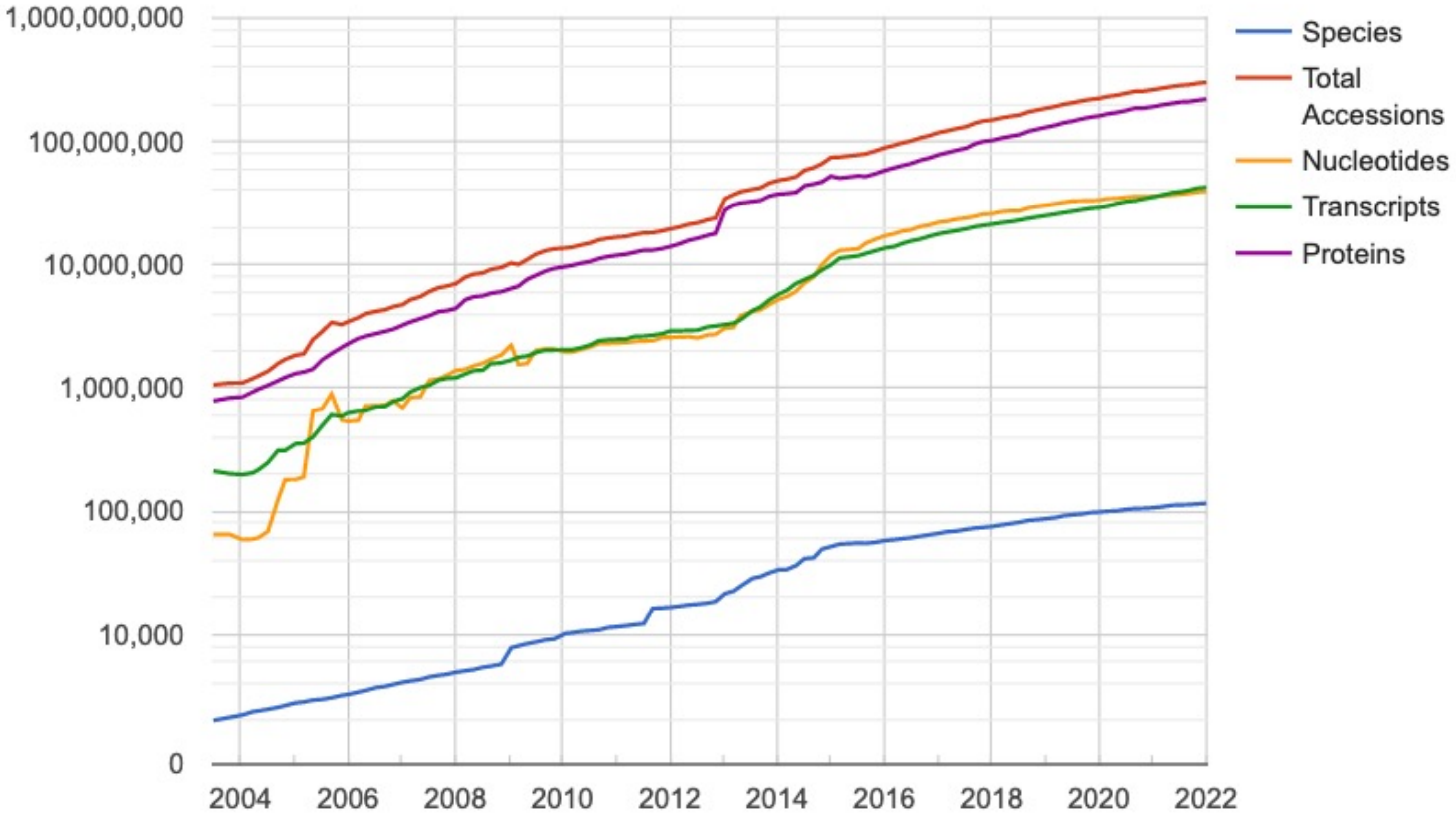
RefSeq

- “The Reference Sequence (RefSeq) collection provides a **comprehensive, integrated, non-redundant, well-annotated set of sequences**, including genomic DNA, transcripts, and proteins”

- “RefSeq transcript and protein records for a subset of organisms, primarily mammals, are **curated by NCBI staff**”

- **Advantages**:
  - Non-redundant
  - Links between nucleic an protein sequences records
  - Updates by NCBI staff and record status indication
  - Data validation and format consistency
  - Overview of information from several records

RefSeq



https://www.ncbi.nlm.nih.gov/refseq/statistics/

Different levels of data curation

- Curation level is indicated in the field ”COMMENT”

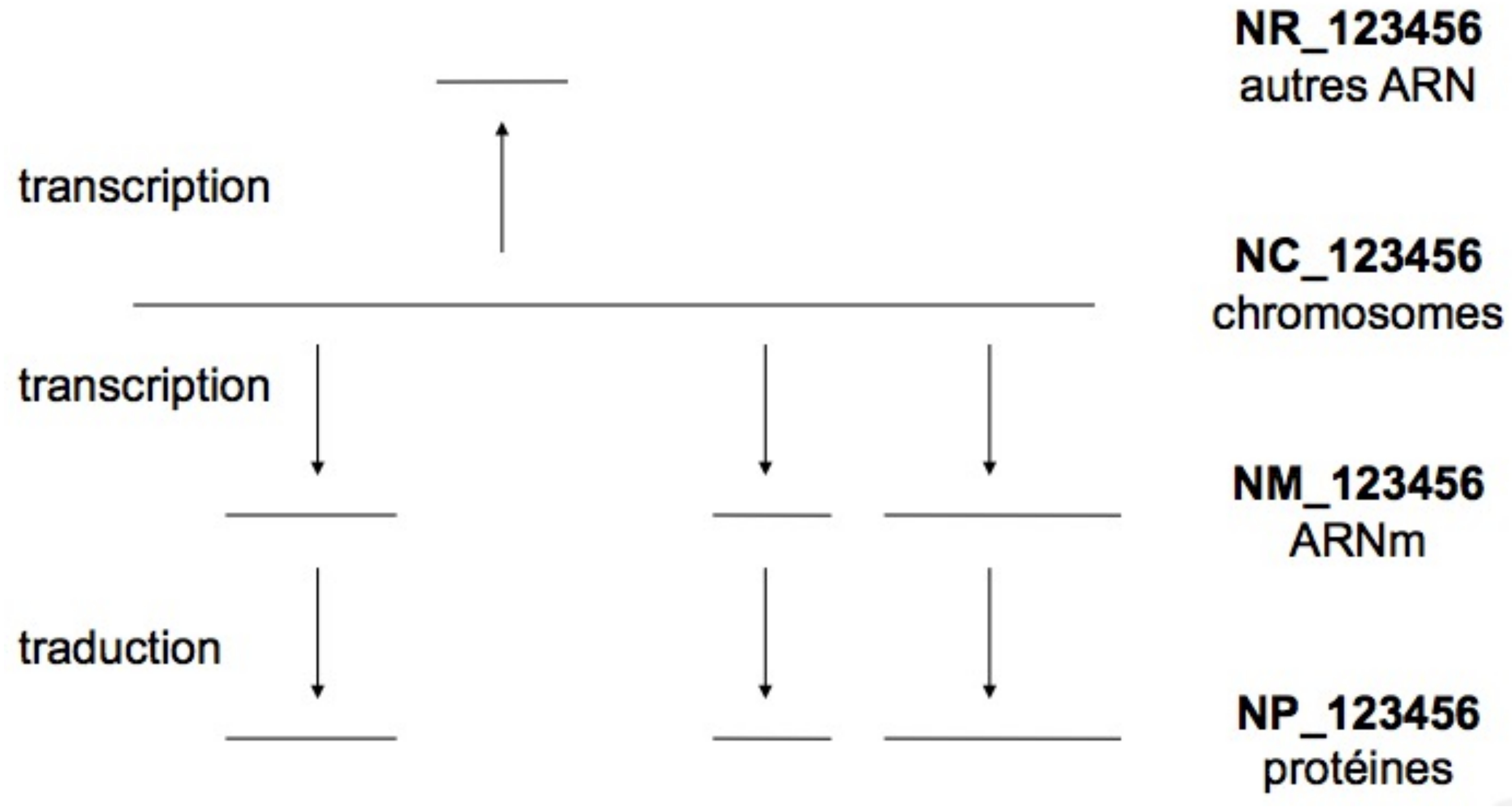
  - **Reviewed**: reviewed by ncbi staff who added information from scientific publications and sequence records

  - **Validated**: NCBI staff have conducted an initial review but the annotation is in progress

  - **Provisional**: not-reviewed, but the record contain probably a true transcript or protein

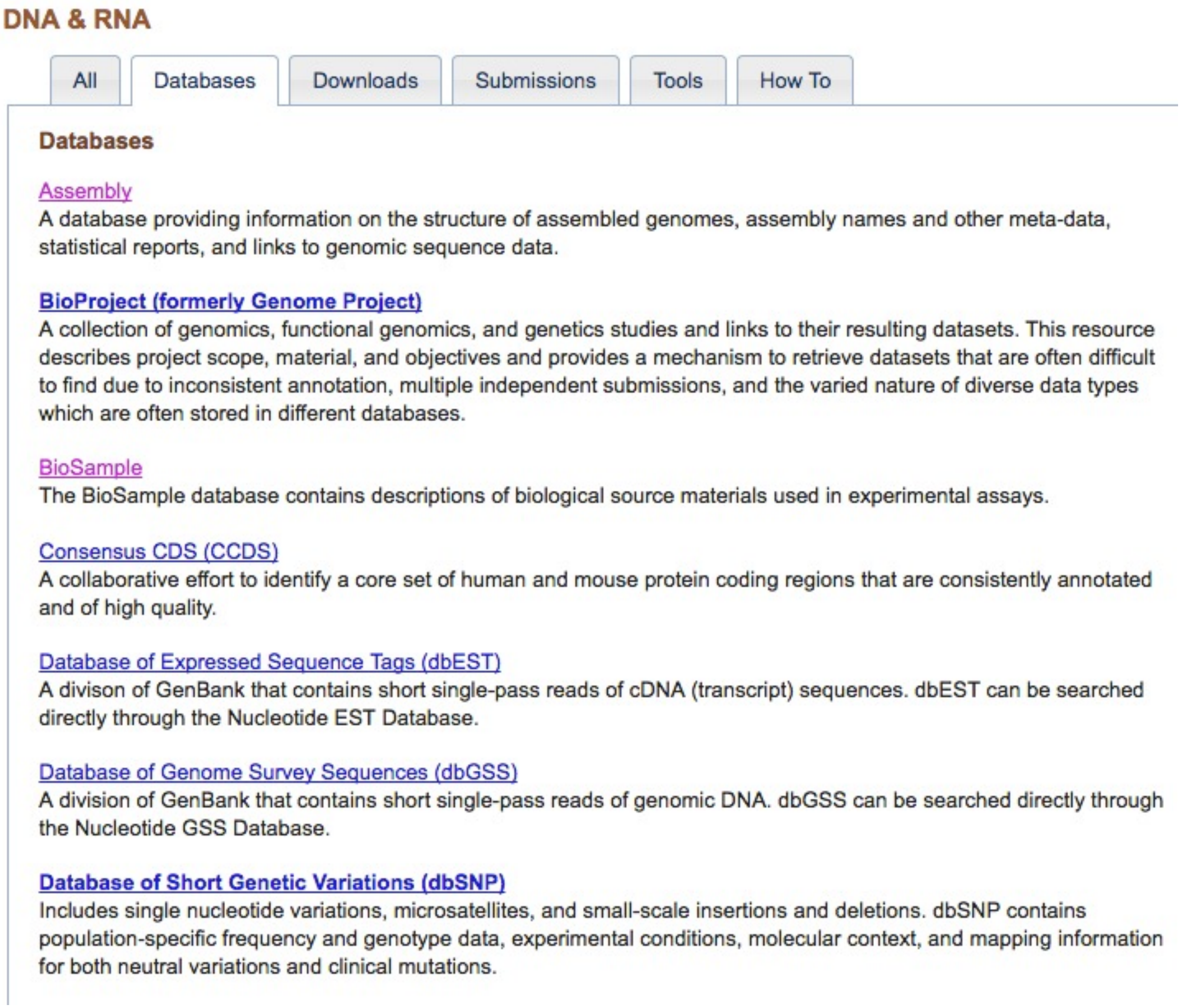
  - **Predicted**: transcript or protein predicted by a software

Links between records



Other NCBI nucleotide databases

http://www.ncbi.nlm.nih.gov/guide/dna-rna/



DNA & RNA

Databases

Assembly
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

BioSample
The BioSample database contains descriptions of biological source materials used in experimental assays.

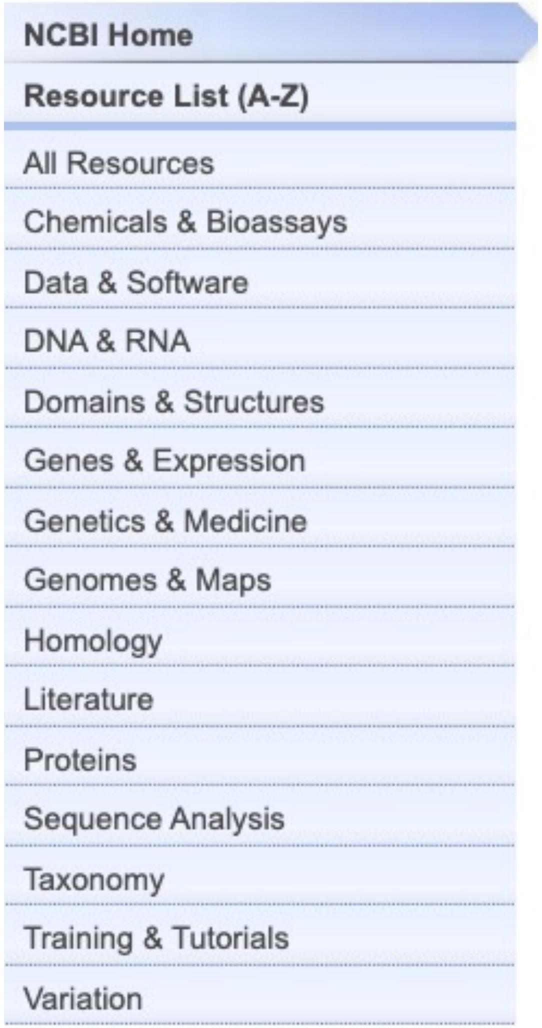
Consensus CDS (CCDS)
A collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality.

Database of Expressed Sequence Tags (dbEST)
A divison of GenBank that contains short single-pass reads of cDNA (transcript) sequences. dbEST can be searched directly through the Nucleotide EST Database.

Database of Genome Survey Sequences (dbGSS)
A division of GenBank that contains short single-pass reads of genomic DNA. dbGSS can be searched directly through the Nucleotide GSS Database.

Database of Short Genetic Variations (dbSNP)
Includes single nucleotide variations, microsatellites, and small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral variations and clinical mutations.

Other NCBI databases



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
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Protein databases

Protein databases

- **Origin of the data**

  - Translation of DNA sequences → a lot of data available in nucleotide database

  - Little protein sequencing because it is long and expensive

- **Stored data**: sequences and annotations

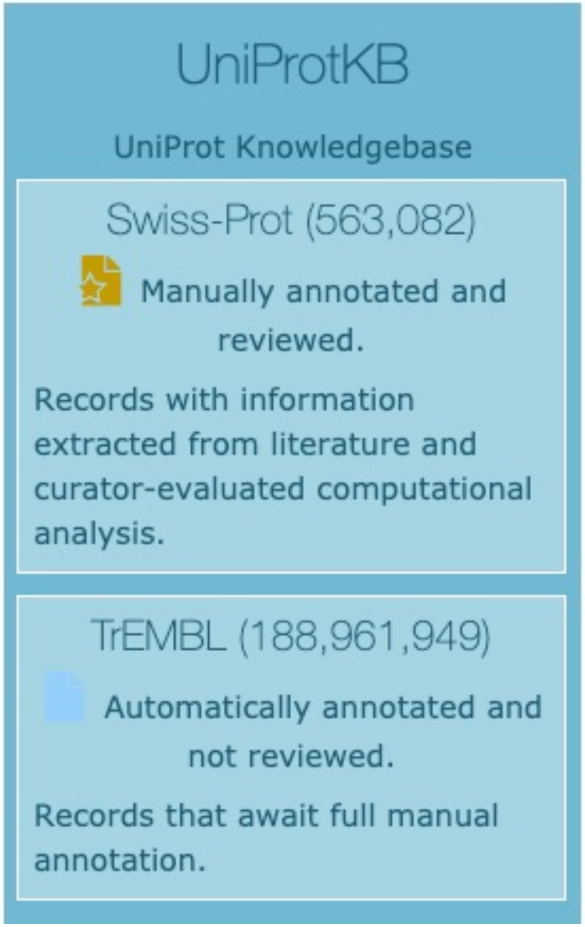
  - Whole proteins

  - Protein fragments

UniProt and its two databases

- **Swiss-Prot**
  - Curated and validated data
  - Highly annotated
  - Low redundancy
  - Many links to other databases

- **TrEMBL**
  - CDS from EMBL automaticaly translated
  - Automatically annotated
  - Additional records to SwissProt that await annotation



UniProtKB

UniProt Knowledgebase

Swiss-Prot (563,082)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (188,961,949)

Automatically annotated and not reviewed.

Records that await full manual annotation.

SwissProt annotations

- Protein **function**(s)
- **Post-translational modifications**
- **Domains** and functional **sites**
- **Secondary structure**
- Participation to a **protein complex**
- **Similarity** to other proteins
- **Conflicts** (uncertain bases)
- **Protein-deficiency diseases**
- **Splicing** variants

SwissProt annotations

- **Source of annotations**
  - Articles dedicated to a protein
  - Review related to a protein family
  - Information from experts
  - Prediction using software (verified by an expert)

- Sources of annotations are **mentioned**

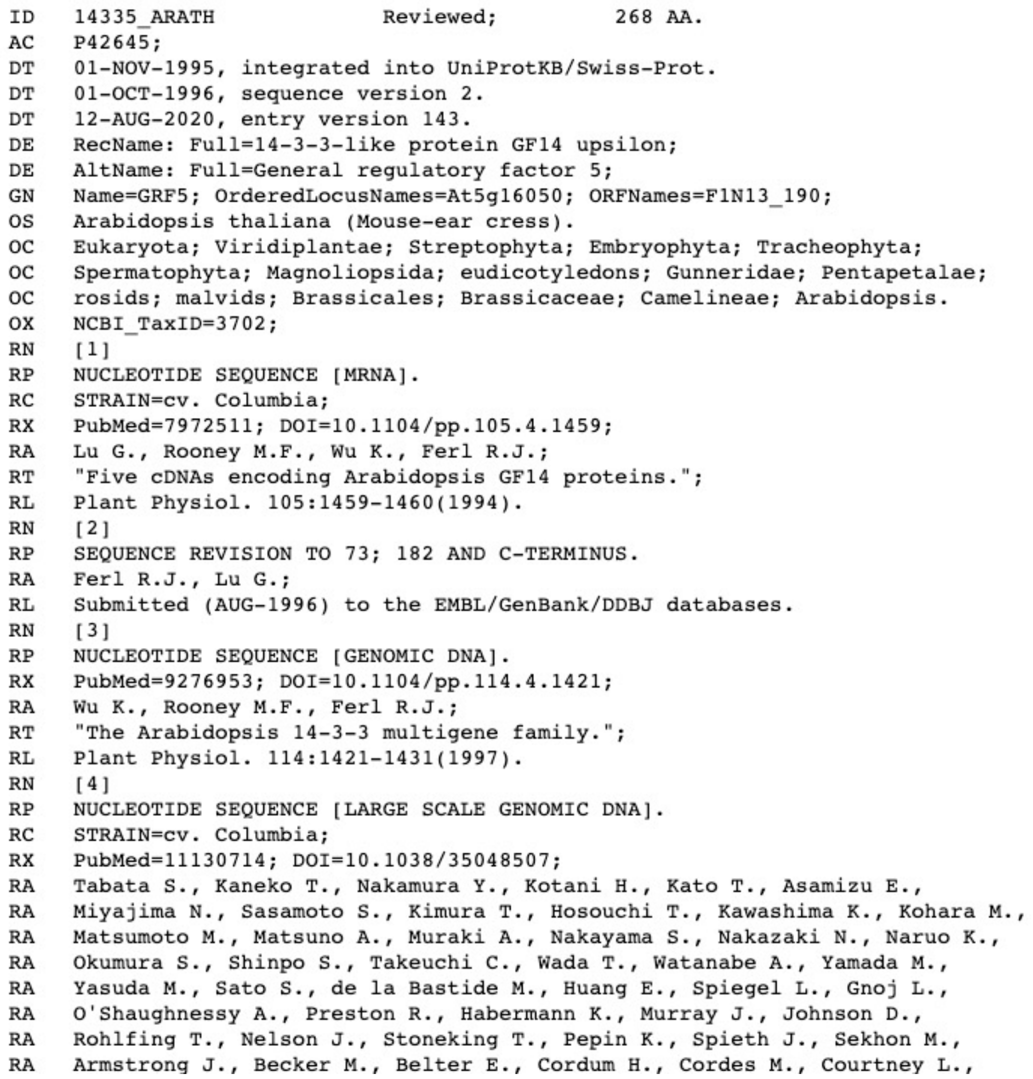
- **Where** in the record?
  - Lines “**CC**” (comments)
  - Lines “**FT**” (localised in the sequence)

UniProt record format

- **Same as EMBL**
  - Each line begins with 2 letters
  - Same keywords
  - But a different format for the features

- **Additional keywords**
  - GN: names of the gene(s) coding the protein
  - OX: links to taxonomy databases
  - CC: comments, highly developed in SwissProt records
  - KW: keywords

UniProt record format

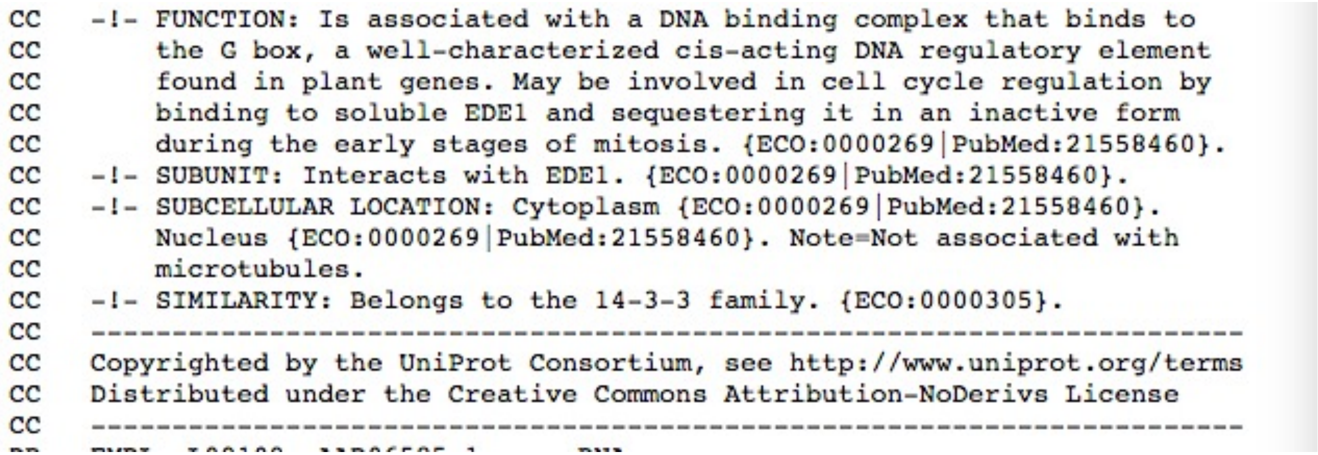


```
ID   14335_ARATH             Reviewed;         268 AA.
AC   P42645;
DT   01-NOV-1995, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 2.
DT   12-AUG-2020, entry version 143.
DE   RecName: Full=14-3-3-like protein GF14 upsilon;
DE   AltName: Full=General regulatory factor 5;
GN   Name=GRF5; OrderedLocusNames=At5g16050; ORFNames=F1N13_190;
OS   Arabidopsis thaliana (Mouse-ear cress).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae;
OC   rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis.
OX   NCBI_TaxID=3702;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   STRAIN=cv. Columbia;
RX   PubMed=7972511; DOI=10.1104/pp.105.4.1459;
RA   Lu G., Rooney M.F., Wu K., Ferl R.J.;
RT   "Five cDNAs encoding Arabidopsis GF14 proteins.";
RL   Plant Physiol. 105:1459-1460(1994).
RN   [2]
RP   SEQUENCE REVISION TO 73; 182 AND C-TERMINUS.
RA   Ferl R.J., Lu G.;
RL   Submitted (AUG-1996) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   PubMed=9276953; DOI=10.1104/pp.114.4.1421;
RA   Wu K., Rooney M.F., Ferl R.J.;
RT   "The Arabidopsis 14-3-3 multigene family.";
RL   Plant Physiol. 114:1421-1431(1997).
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=cv. Columbia;
RX   PubMed=11130714; DOI=10.1038/35048507;
RA   Tabata S., Kaneko T., Nakamura Y., Kotani H., Kato T., Asamizu E.,
RA   Miyajima N., Sasamoto S., Kimura T., Hosouchi T., Kawashima K., Kohara M.,
RA   Matsumoto M., Matsuno A., Muraki A., Nakayama S., Nakazaki N., Naruo K.,
RA   Okumura S., Shinpo S., Takeuchi C., Wada T., Watanabe A., Yamada M.,
RA   Yasuda M., Sato S., de la Bastide M., Huang E., Spiegel L., Gnoj L.,
RA   O'Shaughnessy A., Preston R., Habermann K., Murray J., Johnson D.,
RA   Rohlfing T., Nelson J., Stoneking T., Pepin K., Spieth J., Sekhon M.,
RA   Armstrong J., Becker M., Belter E., Cordum H., Cordes M., Courtney L.,
```

UniProt record format, lines CC

- **Blocks of information** for better readability
  - CC -!- TOPIC: First line of a comment block;
  - CC second and subsequent lines of a comment block

- **Information on many topics**
  - FUNCTION: general description of the function
  - CATALYTIC ACTIVITY: description of reactions catalyzed by enzymes
  - DEVELOPMENTAL STAGE; description of the stages at which the protein is expressed
  - SUBUNIT: complexes of which protein is a part (+partners)
  - …

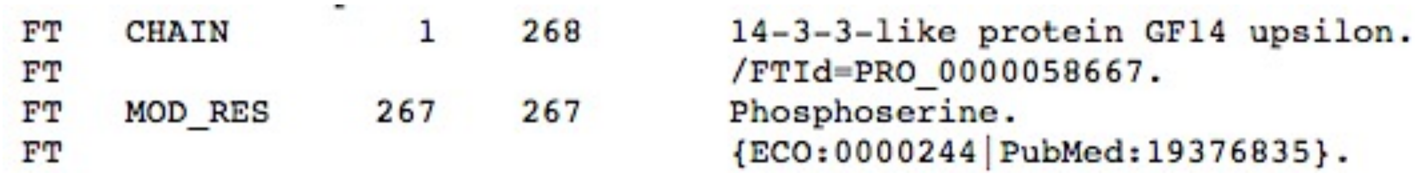


```
CC   -!- FUNCTION: Is associated with a DNA binding complex that binds to
CC       the G box, a well-characterized cis-acting DNA regulatory element
CC       found in plant genes. May be involved in cell cycle regulation by
CC       binding to soluble EDE1 and sequestering it in an inactive form
CC       during the early stages of mitosis. {ECO:0000269|PubMed:21558460}.
CC   -!- SUBUNIT: Interacts with EDE1. {ECO:0000269|PubMed:21558460}.
CC   -!- SUBCELLULAR LOCATION: Cytoplasm {ECO:0000269|PubMed:21558460}.
CC       Nucleus {ECO:0000269|PubMed:21558460}. Note=Not associated with
CC       microtubules.
CC   -!- SIMILARITY: Belongs to the 14-3-3 family. {ECO:0000305}.
CC   -----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   -----------------------------------------------------------------------
```

UniProt record format, lines FT

- **Regions or sites of interest in the sequence**
  - Post-translational modifications
  - Biding sites
  - Active sites (enzymes)
  - Secondary structure
  - Changes in the sequence (including variants)

- **Format in column** (number of characters)
  - 1-2: `FT`
  - 6-13: key
  - 15-20 and 22-27: start and end of object
  - 35-75: description (if necessary on several lines)



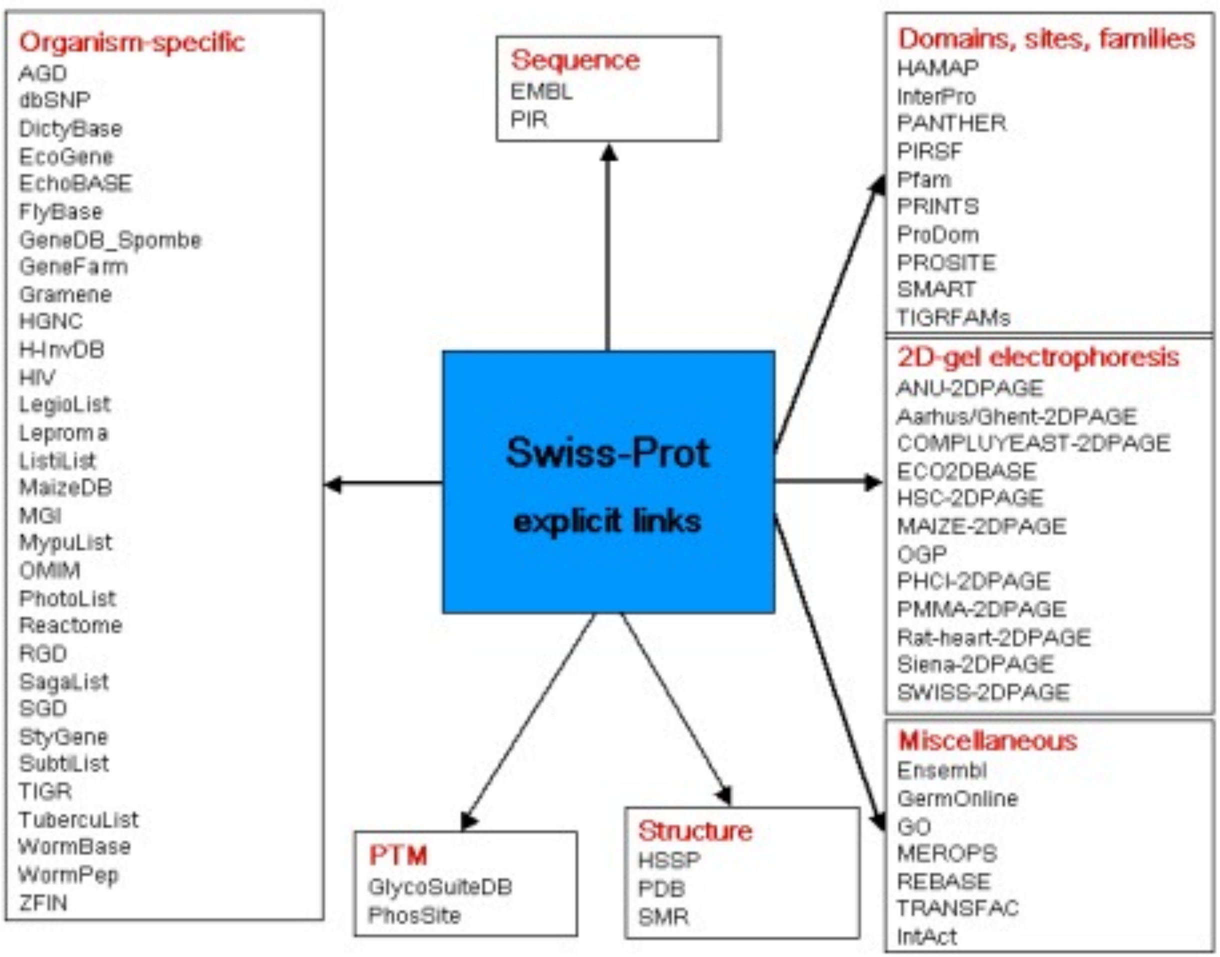
```
FT   CHAIN         1    268       14-3-3-like protein GF14 upsilon.
FT                                /FTId=PRO_0000058667.
FT   MOD_RES     267    267       Phosphoserine.
FT                                {ECO:0000244|PubMed:19376835}.
```

Reliability of information

- Some information are not based on experimental data

- Three keywords to qualify information
  - **Potential**: evidences suggest that the annotation is valid (ex: prediction by a software + contextual consistency)
  - **Probable**: better reliability, early experimental evidence
  - **By similarity**: by sequence similiraty with an experimentally annotated protein

Links to other databases



Swiss-Prot

explicit links

Organism-specific

AGD
dbSNP
DictyBase
EcoGene
EchoBASE
FlyBase
GeneDB_Spombe
GeneFarm
Gramene
HGNC
H-InvDB
HIV
LegioList
Leproma
ListiList
MaizeDB
MGI
MypuList
OMIM
PhotoList
Reactome
RGD
SagaList
SGD
StyGene
SubtiList
TIGR
TubercuList
WormBase
WormPep
ZFIN

Sequence

EMBL
PIR

Domains, sites, families

HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs

2D-gel electrophoresis

ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
ECO2DBASE
HSC-2DPAGE
MAIZE-2DPAGE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE

Miscellaneous

Ensembl
GermOnline
GO
MEROPS
REBASE
TRANSFAC
IntAct

PTM

GlycoSuiteDB
PhosSite

Structure

HSSP
PDB
SMR

Other databases in UniProt

- **UniRef100**: combines identical sequences from any organism
- **UniRef90**: built by clustering UniRef100 sequences that have at least 90% sequence identity
- **UniRef50** : built by clustering UniRef90 seed sequences that have at least 50% sequence identity

- A comprehensive and non-redundant database that contains most of the publicly available protein **sequences** in the world

- UniProt + and other databases (PDB, RefSeq, FlyBase, patents, …)

« Second level » protein databases

- **Starting points**
  - Protein sequences
  - Biological knowledge

- **Sequence analyse** to obtain new data

- Ex: **protein family** databases
  - Clustering proteins with identical or similar functions
  - Generally built by comparing protein sequences
  - HomoloGene, COG, KEGG, SSDB

Motif and protein domain databases

- **A protein family can be characterized by a motif or protein domain**

  - More or less conserved sequence important for the function of the proteins of the family
  - Determined from multiple alignment
  - Several possible representations: consensus sequence, regular expression (regex) , multiple alignment, position weight matrix, hidden Markov model (HMM)

- **Many banks**

  - Prosite, PFAM, Blocks, Prodom, CDD …

Protein knowledge databases

- **Collects various data**
  - Primary data: sequence and annotations
  - Secondary data: from computer predictions (family, motifs, …)

- **Inferring new knowledge**
  - Families built by sequence similarity
  - Pooling of annotations
  - Prediction of the function of unknown proteins

- **Avoids to consult several databases to study a sequence**

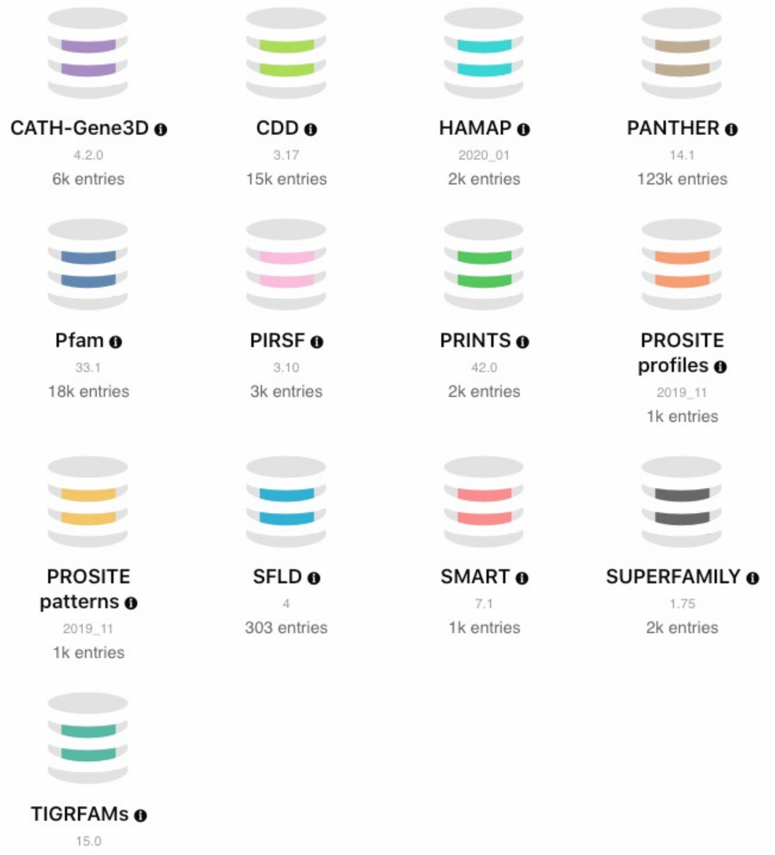
Interpro

- **https://www.ebi.ac.uk/interpro/**

- **Content**
  - superfamilies, families, domains, motifs, functional sites, post-translational modifications, 3D structures

- **Groups data from several databases**
  - Prosite, PFAM, Blocks, Prodom, Smart, Prints, TIGRFams, Superfamily, SCOP, CATH, MSD, …

- **A record**
  - Detailed biological description
  - Representation of the object by the various databases



CATH-Gene3D
4.2.0
6k entries

CDD
3.17
15k entries

HAMAP
2020_01
2k entries

PANTHER
14.1
123k entries

Pfam
33.1
18k entries

PIRSF
3.10
3k entries

PRINTS
42.0
2k entries

PROSITE profiles
2019_11
1k entries

PROSITE patterns
2019_11
1k entries

SFLD
4
303 entries

SMART
7.1
1k entries

SUPERFAMILY
1.75
2k entries

TIGRFAMs
15.0

Protein interaction databases

- **Different levels of interaction**
  - Physical (protein complex)
  - Collaborative (co-processing)

- **Experimental data**
  - yeast two-hybrid, co-immunoprecipitation
  - Biomolecular Interaction Network Data (BIND) database, Database of Interacting Proteins (DIP)

- **Data from scientific articles**
  - Computer analysis of scientific articles, readings by curators
  - Molecular Interaction Database (MINT), IntAct database

- **Development of a data storage format**
  - Standard Initiative (PSI) Molecular Interaction (PSI-MI)

3D structures of proteins

- **1958: first 3D structure** of a protein by Kendrew and Perutz
  - Discovery of the complexity of the 3D structure of a protein

- **Hypothesis**
  - Two proteins with closely related sequences fold similarly
  - Two proteins with close 3D structures have close sequences

- **The 3D structure of proteins is essential for their function**

PDB, 3D structures database

- Worldwide **Protein Data Bank**

- **The only** 3D structure **database** of proteins, amino acids and large biological molecules

- **In 1971,** the Research Collaboratory for Structural Bioinformatics created **PDB**

- **2003, consolidation of the 3 3D structure databases into one**
  - RCSB
  - MSD (Macromolecular Structure Database)
  - PDBJ (Protein Data Bank Japan)

Query Databases

Query from annotations

- **Search for words or expressions** via a query interface

- **What users want**

  - Obtain relevant data: not too many results, but all those related to their problems
  - A user friendly interface
  - Get results quickly
  - Manipulate the data, change format, run calculations

- **Main query system**

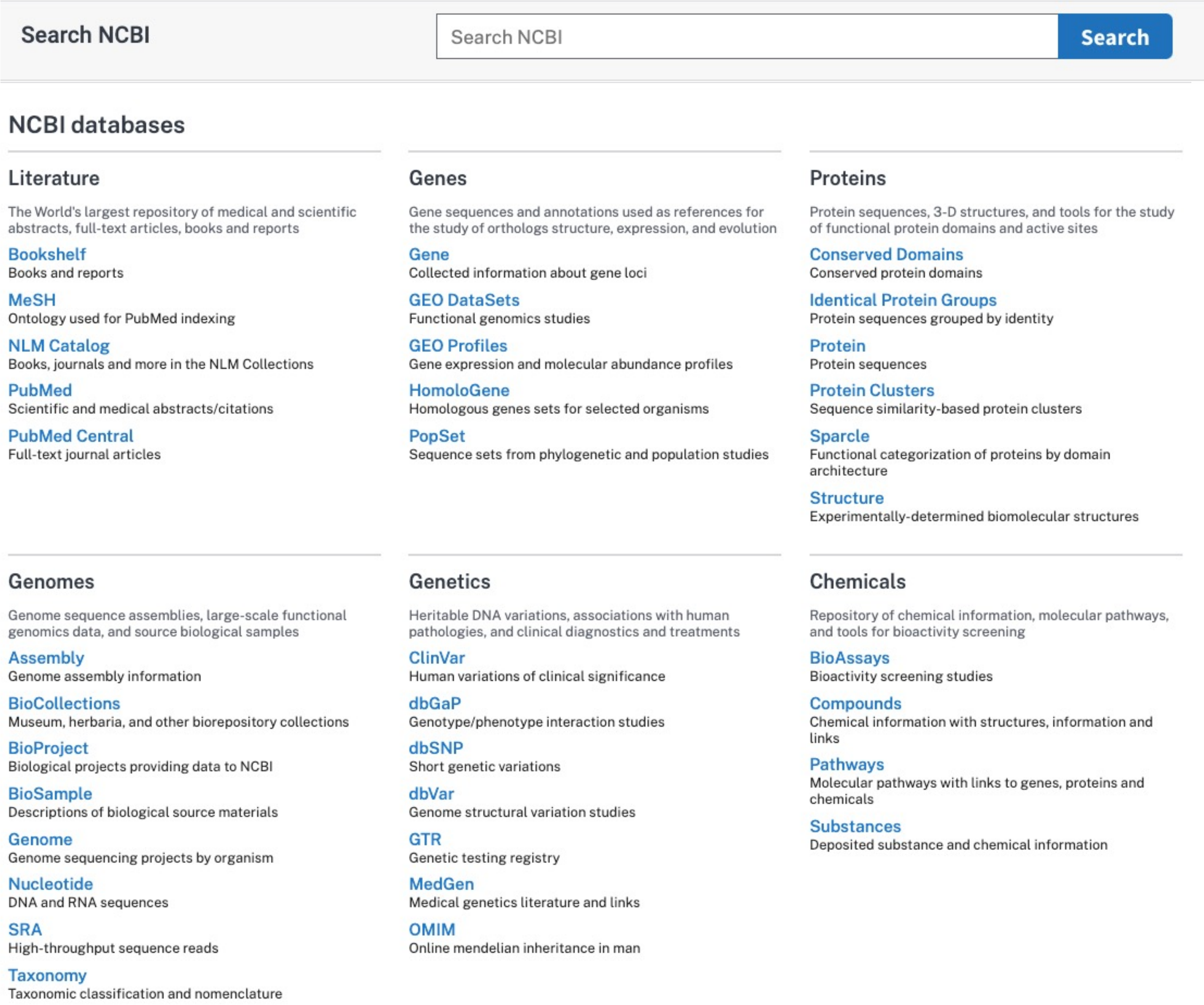
  - EMBL: https://www.ebi.ac.uk/
  - NCBI: https://www.ncbi.nlm.nih.gov/search/

- **Using dedicated tools**:

https://www.ncbi.nlm.nih.gov/books/NBK25501/

```
~ % esearch -db nucleotide -query "M71283" | efetch -format gb
```

Search NCBI



Search NCBI

NCBI databases

Literature

The World's largest repository of medical and scientific abstracts, full-text articles, books and reports

Bookshelf
Books and reports

MeSH
Ontology used for PubMed indexing

NLM Catalog
Books, journals and more in the NLM Collections

PubMed
Scientific and medical abstracts/citations

PubMed Central
Full-text journal articles

Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

Gene
Collected information about gene loci

GEO DataSets
Functional genomics studies

GEO Profiles
Gene expression and molecular abundance profiles

HomoloGene
Homologous genes sets for selected organisms

PopSet
Sequence sets from phylogenetic and population studies

Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

Conserved Domains
Conserved protein domains

Identical Protein Groups
Protein sequences grouped by identity

Protein
Protein sequences

Protein Clusters
Sequence similarity-based protein clusters

Sparcle
Functional categorization of proteins by domain architecture

Structure
Experimentally-determined biomolecular structures

Genomes

Genome sequence assemblies, large-scale functional genomics data, and source biological samples

Assembly
Genome assembly information

BioCollections
Museum, herbaria, and other biorepository collections

BioProject
Biological projects providing data to NCBI

BioSample
Descriptions of biological source materials

Genome
Genome sequencing projects by organism

Nucleotide
DNA and RNA sequences

SRA
High-throughput sequence reads

Taxonomy
Taxonomic classification and nomenclature

Genetics

Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

ClinVar
Human variations of clinical significance

dbGaP
Genotype/phenotype interaction studies

dbSNP
Short genetic variations

dbVar
Genome structural variation studies

GTR
Genetic testing registry

MedGen
Medical genetics literature and links

OMIM
Online mendelian inheritance in man

Chemicals

Repository of chemical information, molecular pathways, and tools for bioactivity screening

BioAssays
Bioactivity screening studies

Compounds
Chemical information with structures, information and links

Pathways
Molecular pathways with links to genes, proteins and chemicals

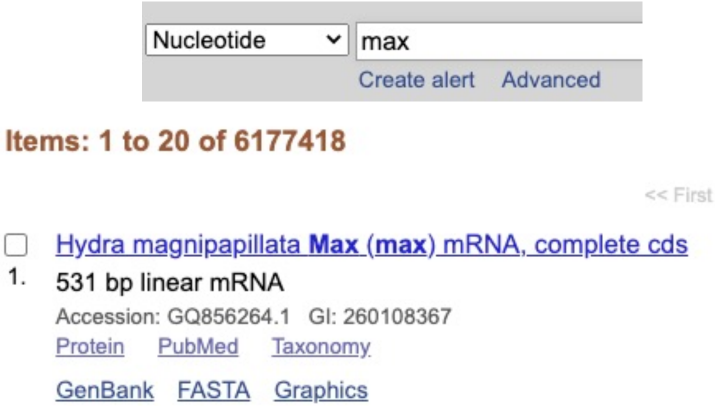
Substances
Deposited substance and chemical information

NCBI, word search

- **Which entries in the nucleic data bank contain the MAX gene?**

  - Enter `max` in the query field
    → Search for the word “max” throughout the text of the records
    → 6,177,418 records

  - Enter now `max [gene]`
    → Search for the word "max" in the fields corresponding to the gene name
    → Targeted search: 2,100 records

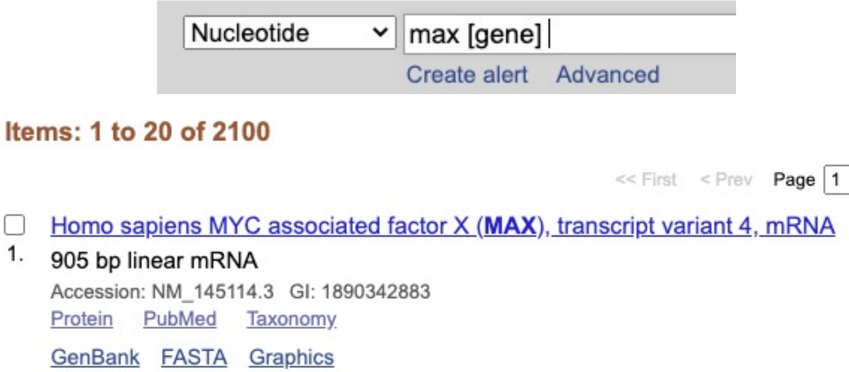


Nucleotide max
Create alert Advanced

Items: 1 to 20 of 6177418

<< First

□ 1. Hydra magnipapillata **Max** (**max**) mRNA, complete cds
531 bp linear mRNA
Accession: GQ856264.1 GI: 260108367
Protein PubMed Taxonomy
GenBank FASTA Graphics



Nucleotide max [gene]
Create alert Advanced

Items: 1 to 20 of 2100

<< First < Prev Page 1

□ 1. Homo sapiens MYC associated factor X (**MAX**), transcript variant 4, mRNA
905 bp linear mRNA
Accession: NM_145114.3 GI: 1890342883
Protein PubMed Taxonomy
GenBank FASTA Graphics

NCBI, use search fields

- **Field**: information listed in a specific part of the entry
  - All the values of a field are listed in an index

- Allows **better targeting of queries**
  - Words are not searched throughout the text of the entry

- **Syntaxe**: `searched_word [field]`

- **Field examples** (depending on the database):
  - `[gene]`: gene name
  Be careful because not all the authors put the name of the gene in the right place !
  - `[protein]`: protein name
  Same remark as above
  - `[organism]`: species name or other taxonomic level

NCBI, combination of terms

- Three **boolean operators***: **AND, OR, NOT**

- Ex: In the nucleotide bank:

```
  - rattus norvegicus [organism] AND mus musculus
[organism]
```

→ 1 record: "Synthetic construct chimeric tyrosine hydroxylase."

```
  - rattus norvegicus [organism] OR mus musculus
[organism]
```

→ 2,063,974 entries

The sequence comes from either the rat or the mouse...

```
  - rattus norvegicus [organism] NOT mus musculus
[organism]
```

→ 334,078 entries

All the rat sequences except the chimeric sequence

* Computer search tools to sort the results of a query more precisely

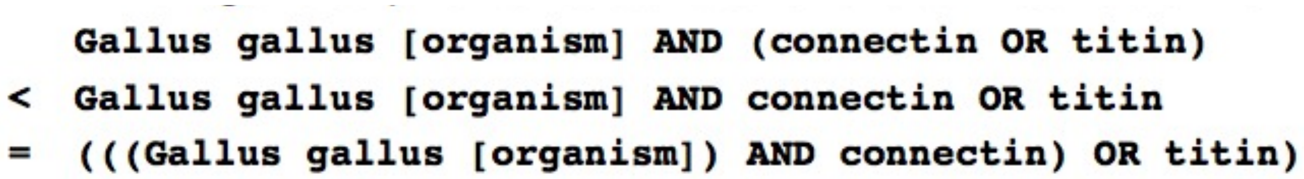
NCBI, how to build a query

- **Determine search criteria**
  - Do not forget declinations (plural, gender, ...)
  - Do not forget synonyms

- **Use fields to limit search**

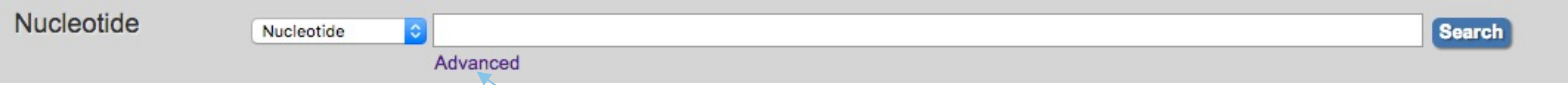
- **Combine these criteria with the right operators**
  - If different criteria complement each other: `AND`
  When querying several fields
  - If alternative between several terms: `OR`
  When several terms for a same field

- **Use brackets to determine priorities**
  - Ex:

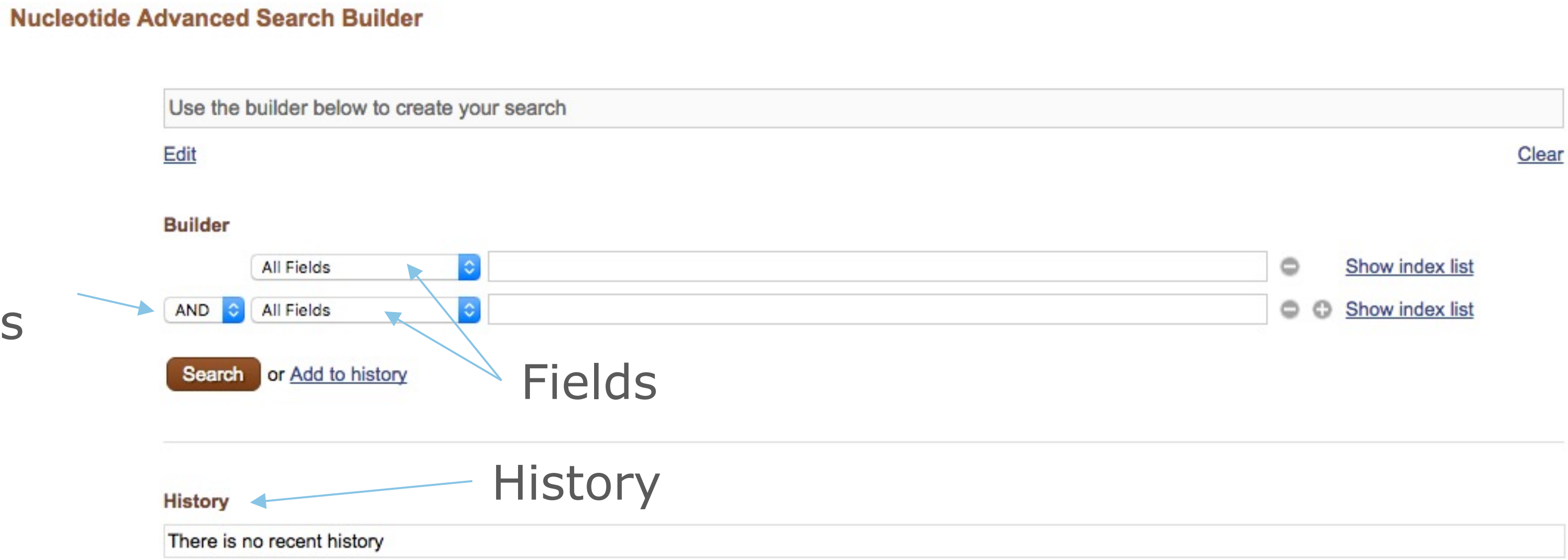


```
   Gallus gallus [organism] AND (connectin OR titin)
<  Gallus gallus [organism] AND connectin OR titin
=  (((Gallus gallus [organism]) AND connectin) OR titin)
```

NCBI, advanced search



Advanced search



Use the builder below to create your search

There is no recent history

Boolean operators

Fields

History

Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille - Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | http://eep.univ-lille.fr/
Tél. +33 (0)3 20 43 40 16