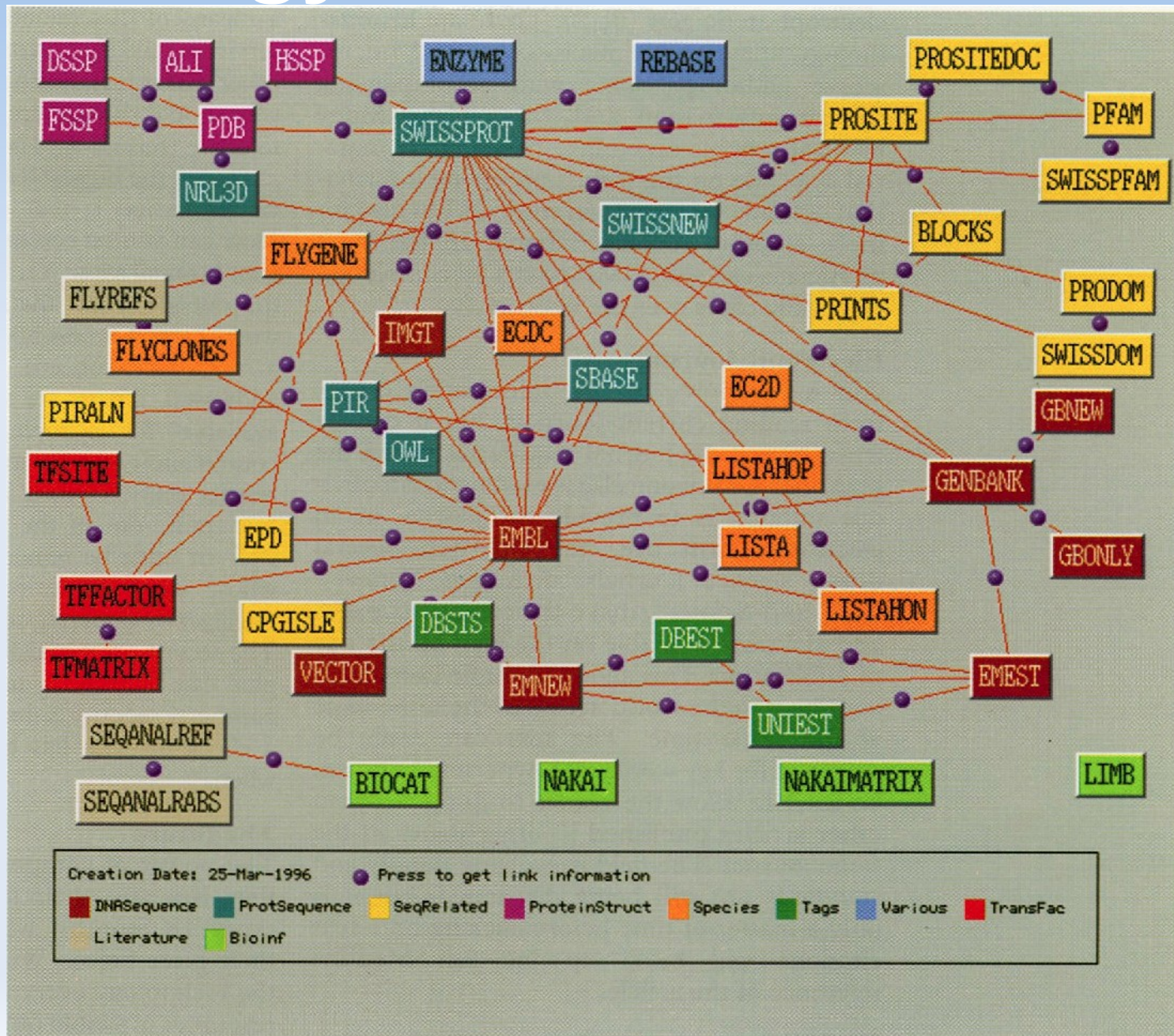


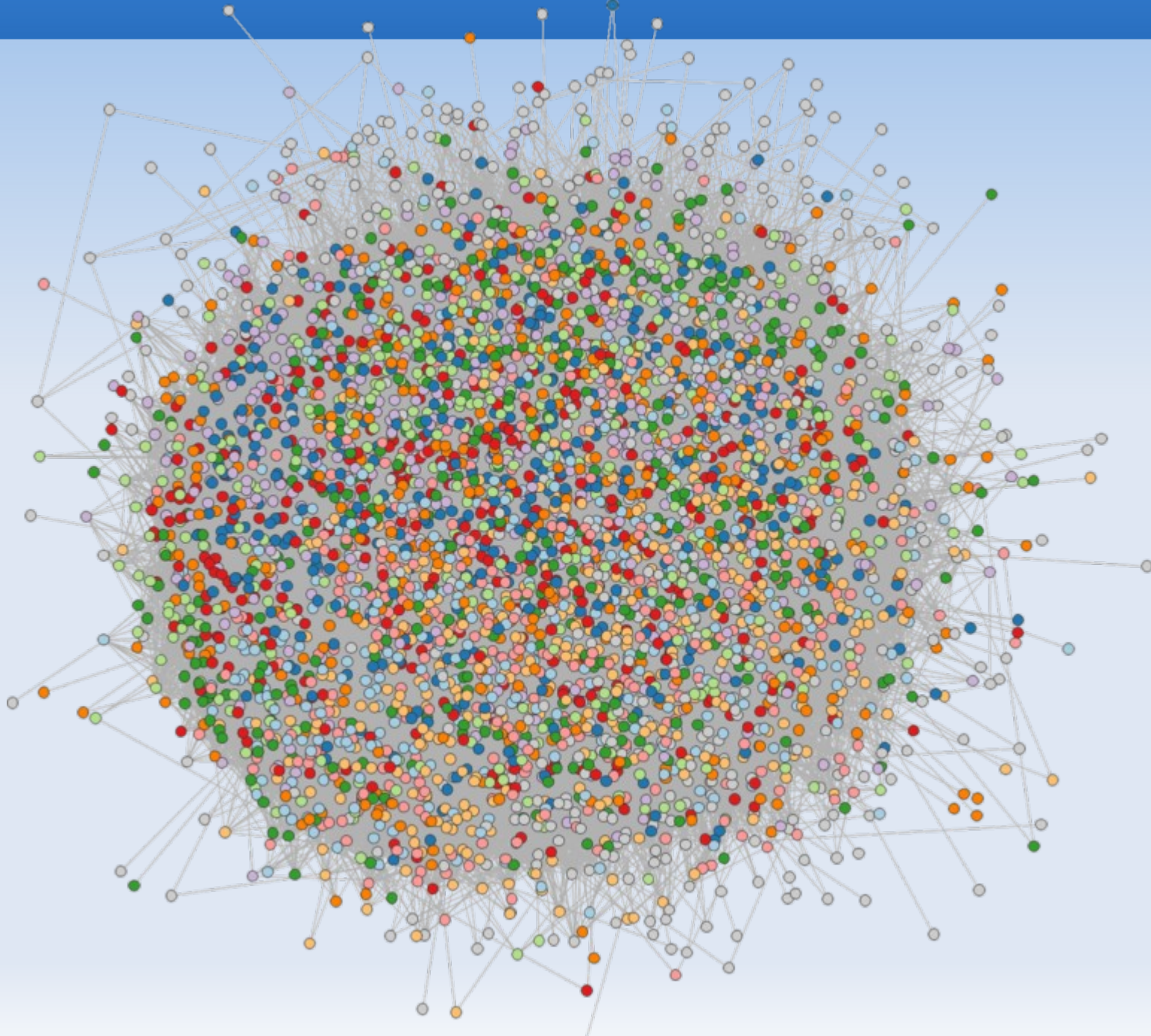
Molecular biology databases

Eszter Ari
ELTE, Dept. Genetics
arieszter@gmail.com

Nr. and connections of molecular biology dbs. Back in 1996



And nowadays...



Publications about databases

Nucleic Acids Research „Database Issue” (every Jan)

Published online 30 November 2012

Nucleic Acids Research, 2013, Vol. 41, Database issue **DI-D7**
doi:10.1093/nar/gks1297

The 2013 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection

Xosé M. Fernández-Suárez^{1,*} and Michael Y. Galperin^{2,*}

¹Cambridge, CB24 6DZ, UK and ²National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received November 14, 2012; Accepted November 15, 2012

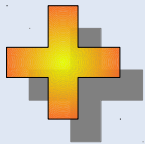
ABSTRACT

The 20th annual Database Issue of *Nucleic Acids Research* includes 176 articles, half of which describe new online molecular biology databases

NEW AND UPDATED DATABASES

This 1300-page virtual volume represents the 20th annual Database Issue of *Nucleic Acids Research* (NAR). It includes descriptions of 88 new online databases, 77 update articles on databases that have been previously

- *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*



DATABASE The Journal of Biological
Databases and Curation

Database column: in *Bioinformatics*,
BMC Bioinformatics journals

Different types of traditional molecular biology databases

1.

Primary databases

- Nucleotide sequence dbs
- Other: i.e. protein structure dbs ← X-ray, NMR

2.

Secondary or derived databases

- Protein sequence dbs ← translated from cDNS
- Motif dbs (i.e: promoters)

3.

Tertiary: network databases

- Connections of components of primary and secondary dbs

Other databases

- Genome, taxonomic, publications, ...

The major nucleotide sequence databases

- **EMBL** - European Molecular Biology Laboratory: „European Nucleotide Archive”
 - **EBI** (European Bioinformatics Institute) maintain it
 - founded in 1980 Heidelberg (D)
 - today: Hinxton (UK)
 - <http://www.ebi.ac.uk/ena/>
- **GenBank** or **Nucleotide**
 - **NCBI** (National Center for Biotechnology Information)
 - founded in 1979 Los Alamos (New Mexico, USA)
 - since 1992 Bethesda, Maryland
 - <http://www.ncbi.nlm.nih.gov/nucleotide/>
- **DDBJ** - DNA Database of Japan
 - **CIB** - Center for Information Biology, Mishima, Japan
 - www.ddbj.nig.ac.jp



Data exchange and synchronization between data warehouses

- INSDC: International Nucleotide Sequence Database Collaboration

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

What is a database?

- Same quality of data
- Structured data
- Stored on computer
- Searchable
- Sortable
- Editable



What is a data source?

- Online accessible
- Free (Open data)
- Community used
- Committee droved
- But in biology...
- we still name these as databases

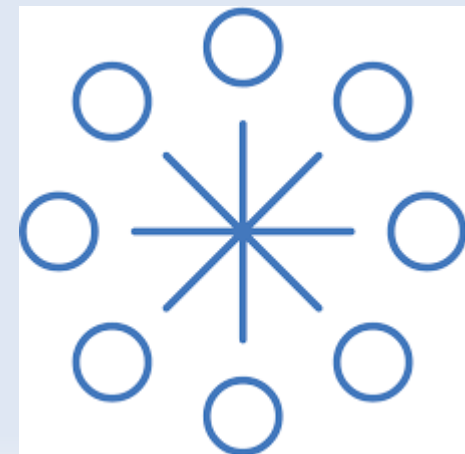


Data storing and seaching

- Information system:
 - NCBI Entrez (renamed to Search), Google, EBI
- Search engine:
 - Blast, SQL, web service
- Storing system:
 - Database system
- Data:
 - Sequence, flat file, table, text

Identifier or ID

- Name is NOT specific
 - i.e: SMAD2 = hMAD-2, JV18-1, MADR2, MADH2, SMAD family member 2, Mad-related protein 2, Mothers against decapentaplegic homolog 2, MAD homolog 2, Mothers against DPP homolog 2, Receptor-regulated SMAD, R-SMAD
- ID indicates an entity in the database
 - i.e. SMAD2:
 - Uniprot: Q15796
 - Ensembl: ENSG00000175387
 - NCBI Gene: 4087
- Translate between Dbs
 - Mapping



Information about a nucleotide or protein sequence

More or less - depending on the authors and the database standards

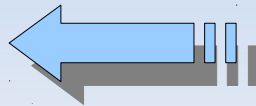
- **Sequence**
- **Genomic information:** location on the chromosome, location of introns, UTRs, regulatory regions, etc.
- **Structural information:** protein structure, fold type, etc.
- **Gene expression:** in different tissues, developmental stages, phenotypes, diseases, etc.
- **Evolution:** homologs, taxonomic distribution, allele frequencies, etc.
- **Functions:** molecular function, role in a pathways, role in diseases, etc.

Structure of an NCBI nucleotide record/entry

- Table (pl. GenBank)

- Record X

- **Annotation**

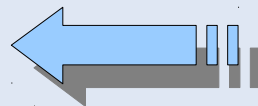


word based searching

- **Field 1** (i.e. Locus)
- **Field 2** (i.e. Definition)
- Etc.
- **Feature table**

- Sequence features: start, end, exons, introns, etc

- **Sequence**



similarity searching (BLAST)

- **Field n** (i.e. cgagcatgcatctagtagcagcgactac)

An NCBI nucleotide entry

Homo sapiens cytochrome c, somatic (CYCS), mRNA

NCBI Reference Sequence: NM_018947.6

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS	NM_018947	5432 bp	mRNA	linear	PRI 31-AUG-2019
DEFINITION	Homo sapiens cytochrome c, somatic (CYCS), mRNA.				
ACCESSION	NM_018947				
VERSION	NM_018947.6				
KEYWORDS	RefSeq; RefSeq Select.				
SOURCE	Homo sapiens (human)				
ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 5432)				
AUTHORS	Neubauer K, Wozniak-Stolarska B and Krzystek-Korpacka M.				
TITLE	Peripheral Lymphocytes of Patients with Inflammatory Bowel Disease Have Altered Concentrations of Key Apoptosis Players: Preliminary Results				
JOURNAL	Biomed Res Int 2018, 4961753 (2018)				
PUBMED	30515402				

FEATURES	Location/Qualifiers
<u>source</u>	1..5432 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon: 9606 " /chromosome="7" /map="7p15.3"
<u>gene</u>	1..5432 /gene="CYCS" /gene_synonym="CYC; HCS; THC4" /note="cytochrome c, somatic" /db_xref="GeneID: 54205 " /db_xref="HGNC: HGNC:19986 " /db_xref="MIM: 123970 "
<u>exon</u>	1..61 /gene="CYCS" /gene_synonym="CYC; HCS; THC4" /inference="alignment:Splign:2.1.0"
<u>exon</u>	62..238 /gene="CYCS"

ORIGIN

```
1 agagagtggg gacgtccggc ttcggagcgg gagtgttcgt tgtgccagcg actaaaaaga
61 gaattaaata tgggtgatgt tgagaaaggc aagaagattt ttattatgaa gtgttcccag
121 tgccacaccg ttgaaaaggg aggcaagcac aagactgggc caaatctcca tggctctctt
181 gggcggaaga caggtcaggc ccctggatac tcttacacag ccgccaataa gaacaaaggc
241 atcatctggg gagaggatac actgatggag tatttggaga atcccaagaa gtacatccct
301 ggaacaaaaa tgatctttgt cggcattaag aagaaggaag aaagggcaga cttaatagct
361 tatctcaaaa aagctactaa tgagtaataa ttggccactg ccttatttat taaaaacag
421 aatgtctca tgactttttt atgtgtacca tcctttaata gatctcatal accagaattc
481 agatcatgaa tgactgacag aatattttgt tgggcagtcc tgatttaaaa ctaagactgg
541 cttgtgggta aatgaatatg ttcagttttt gaattttaat agtaactcca attcagtaaa
601 tggtatcact gtttaccctt tttaaagata tgattagact tcgttagtaa tgttcaactt
661 ttcacaaaga tgggtgagtg catcttaaaa cttactggag attgggtttta tatttagatt
721 tatataactg gttatgtgaa tatatttaaa tactggggaa attgcttcac tgtcttagaa
781 ccaagcaaga ttcacctgtg ttttgtgttc atgttcattt gcctcttaaa ggcaagggtt
841 gaagataaat aaggtagcaa tgtctatagt tttggcctta actatgccaa tctaattata
```


Submit a sequence to one of the primary nucleotide database

- From different research groups, genome sequencing programs
 - *sequence submission* → **unique accession number** to every sequence entry
 - it could be any kind of annotated sequence: gene, chromosome, redundant, partial or non-coding sequence as well
 - sequence submission to one of the 3 main databases is required to publish the results in sci. journals



Reliability

- It is important to know that a record contains information that the authors found important to give (beside some obligatory parts)
- Therefore sometimes a record is not up to date or contains some incorrect informations
- Double check everything using other resources as well!
- Use reference sequences (**RefSeq**)!



© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

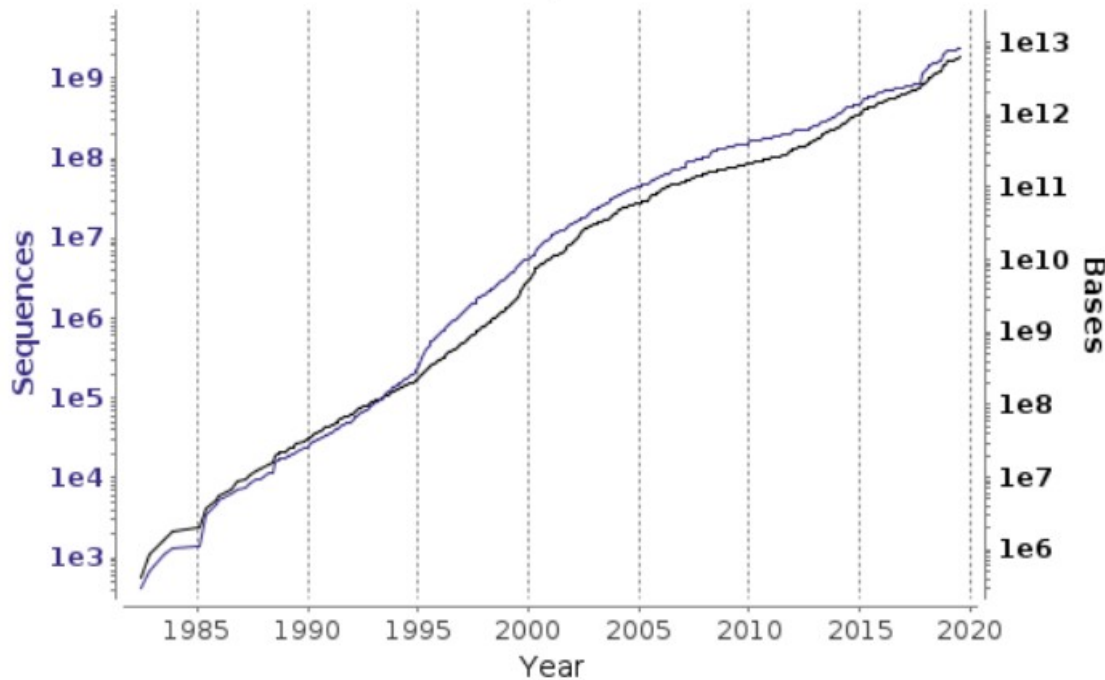


"Well, it certainly looks like your DNA. How many times have I told you to wear gloves before touching anything?"

Size of EMBL ENA

Assembled/annotated sequence growth

09-Sep-2019



— Sequences (2,373.6 millions) — Bases (6,223.4 billions)

Year	Million (mega) sequence record	Billion (giga) bases
2019	2,374	6,223
2016	759	1,855
2013	327	689
2011	199	301

Non redundant databases

- NCBI RefSeq



- <http://www.ncbi.nlm.nih.gov/refseq/>
- extensive, integrated, well annotated
- genomic DNA, cDNA, protein
- Reference genome sequences

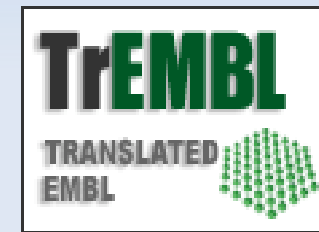
Protein sequence databases: History

■ Swiss-Prot



- Since 1986
- It was maintained by SIB (Swiss Institute of Bioinformatics) and EBI
- Best annotated database (annotations by hand)
- → it was integrated to *UniProt*

■ TrEMBL



- Translated EMBL, automated sequence translations and annotations
- → it was integrated to *UniProt*

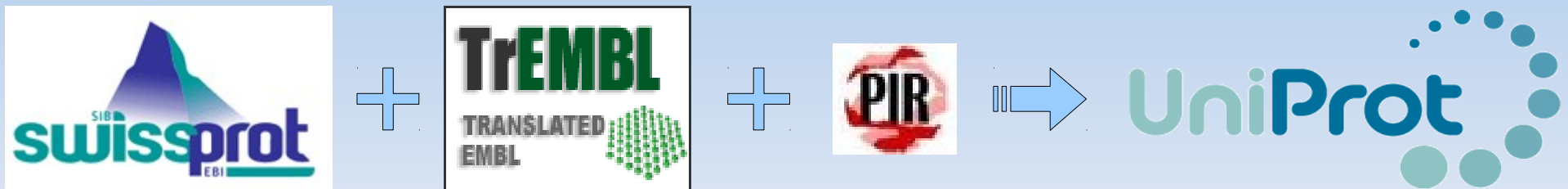
■ PIR (Protein Identification Resource) <http://pir.georgetown.edu/>

- It was founded in 1960s by *Margaret Dayhoff* and the National Biomedical Research Foundation (USA)
- annotations by hand
- → it was integrated to *UniProt*





Protein databases

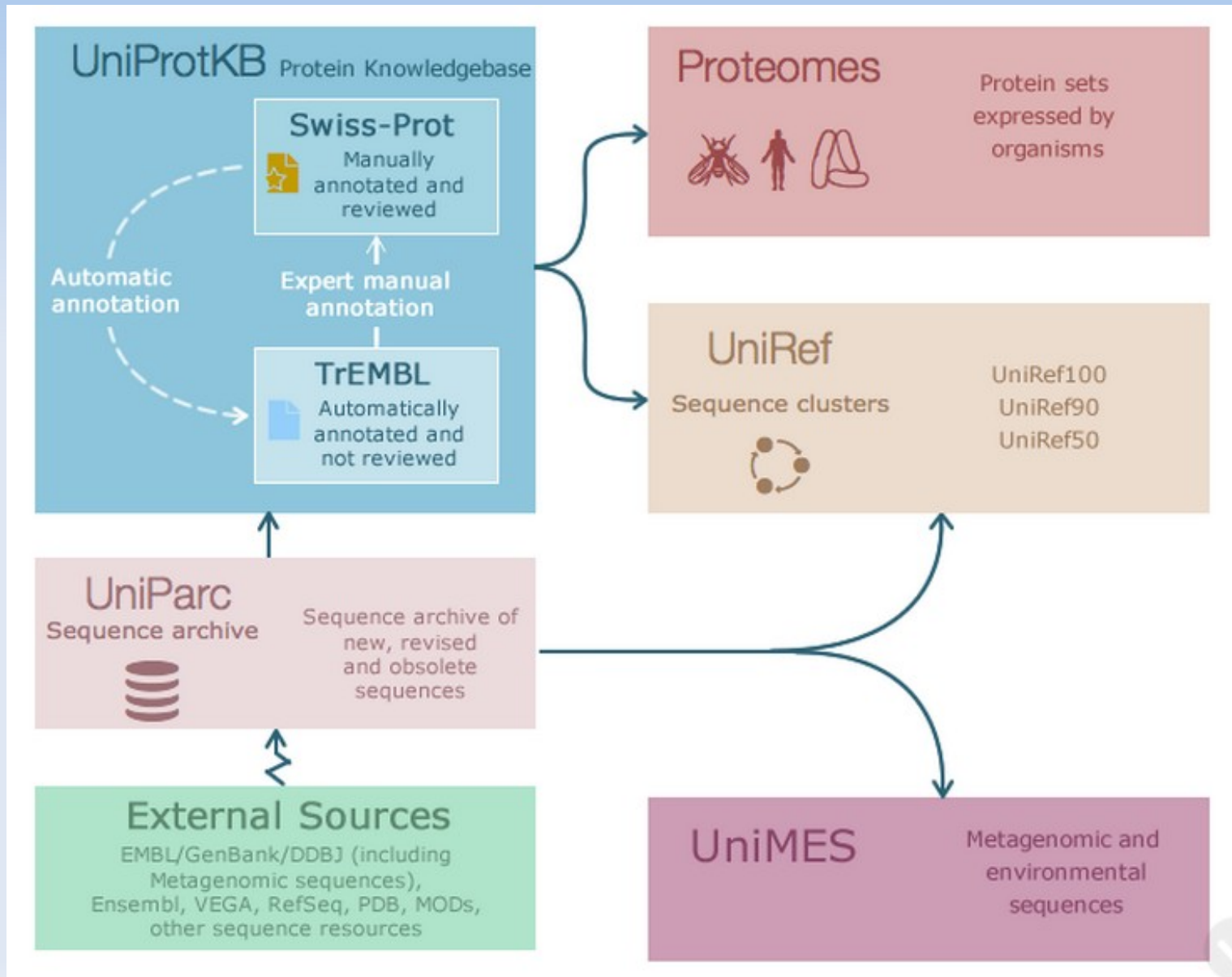
- **UniProt:** Universal Protein Resource: <http://www.uniprot.org>



UniProt Consortium (2002)

- three layers of the database :
 - **UniProtKB:** *UniProt Knowledgebase*, well annotated protein database
 - 2 parts:
 -  Reviewed: manually annotated (Swiss-Prot), 560,000 protein sequences
 -  Unreviewed: automatically annotated (TrEMBL), 168,000,000 p. seq.s
 - **UniRef:** *UniProt Reference Clusters*, protein sequence clusters → speeds up sequence similarity searches (BLAST)
 - **UniParc:** *UniProt Archive*, a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

UniProt



A UniProt entry in text format

```
ID      RBL_WHEAT              Reviewed;              477 AA.
AC      P11383; Q7YKX2;
DT      01-JUL-1989, integrated into UniProtKB/Swiss-Prot.
DT      01-AUG-1990, sequence version 2.
DT      31-JUL-2019, entry version 136.
DE      RecName: Full=Ribulose biphosphate carboxylase large chain;
DE              Short=RuBisCO large subunit;
DE              EC=4.1.1.39;
DE      Flags: Precursor;
GN      Name=rbcL;
OS      Triticum aestivum (Wheat).
OG      Plastid; Chloroplast.
OC      Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC      Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP clade;
OC      Pooideae; Triticeae; Triticum.
OX      NCBI_TaxID=4565;
RN      [1]
RP      NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RA      Terachi T., Ogiwara Y., Tsunewaki K.;
RT      "The molecular basis of genetic diversity among cytoplasms of Triticum
RT      and Aegilops. VI. Complete nucleotide sequences of the rbcL genes
RT      encoding H- and L-type rubisco large subunits in common Wheat and Ae.
RT      crassa 4x.";
RL      Jpn. J. Genet. 62:375-387(1987).
RN      [2]
```

...

The same UniProt entry on the website

UniProtKB - P11383 (RBL_WHEAT)

Basket

BLAST Align Format Add to basket History

Other tutorials and videos Help video Feedback

Display

Entry

Publications

Feature viewer

Feature table

None

Function

Names & Taxonomy

Subcell. location

Pathol./Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Protein | **Ribulose biphosphate carboxylase large chain**

Gene | **rbcl**

Organism | *Triticum aestivum (Wheat)*

Status

Reviewed - Annotation score: ●●●●●● - Experimental evidence at protein levelⁱ

Functionⁱ

RuBisCO catalyzes two reactions: the carboxylation of D-ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate in the photorespiration process (PubMed:2928307). Both reactions occur simultaneously and in competition at the same active site. 1 Publication

Miscellaneous

The basic functional RuBisCO is composed of a large chain homodimer in a "head-to-tail" conformation. In form I RuBisCO this homodimer is arranged in a barrel-like tetramer with the small subunits forming a tetrameric "cap" on each end of the "barrel".

Catalytic activityⁱ

Genome browsers

- Ensembl:

- <http://www.ensembl.org>
- Maintained by EBI and Sanger Inst.



- NCBI Genome Data Viewer:

- <https://www.ncbi.nlm.nih.gov/genome/gdv/>

Genome Data Viewer

- UCSC Genome Browser:

- <http://genome.ucsc.edu>
- University of California



Protein structure

- PDB - Protein Data Bank
 - <https://www.rcsb.org/>
 - 3D structures of molecules
 - ~144,000 protein structures (and a few structures of DNA and RNA molecules)



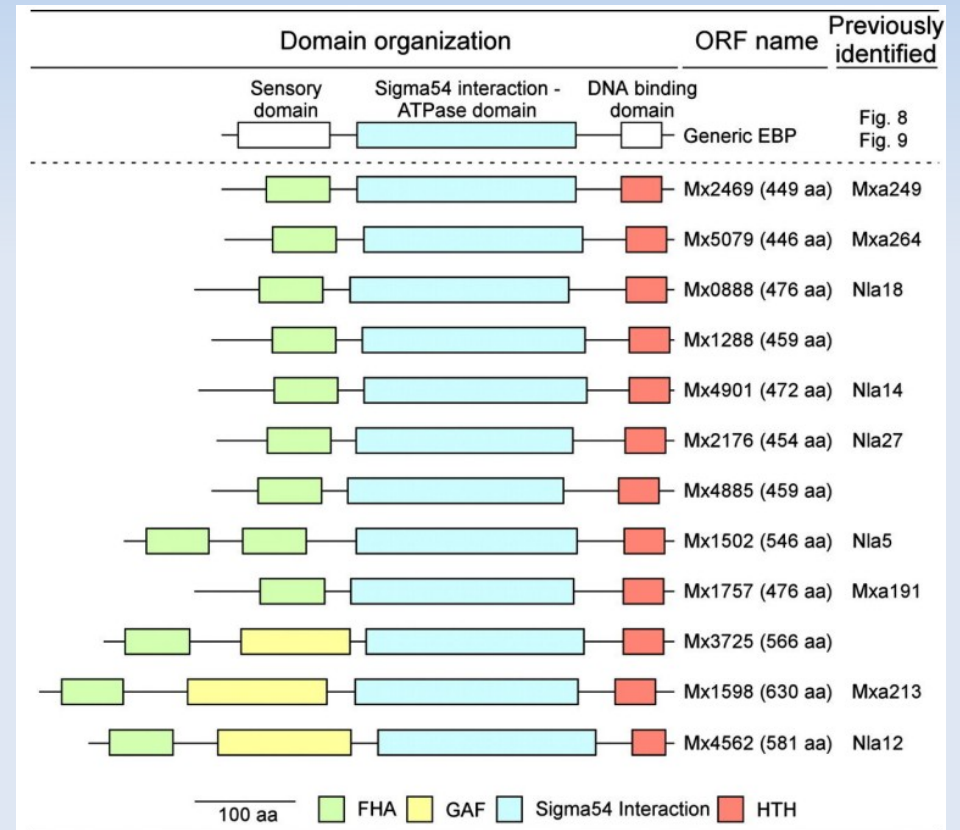
Below the protein

- Domain, families:

- Pfam
- InterPro
- PROSITE

- Motif

- ELM
- Phosphosite



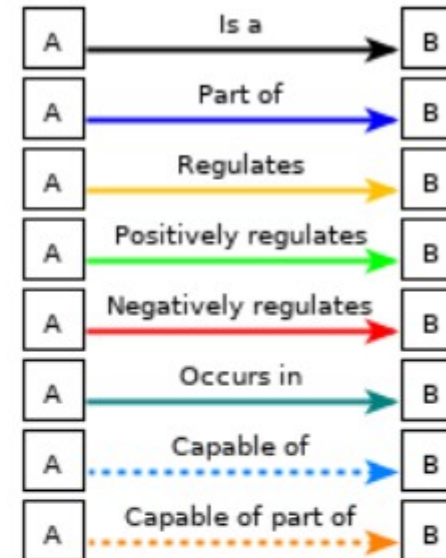
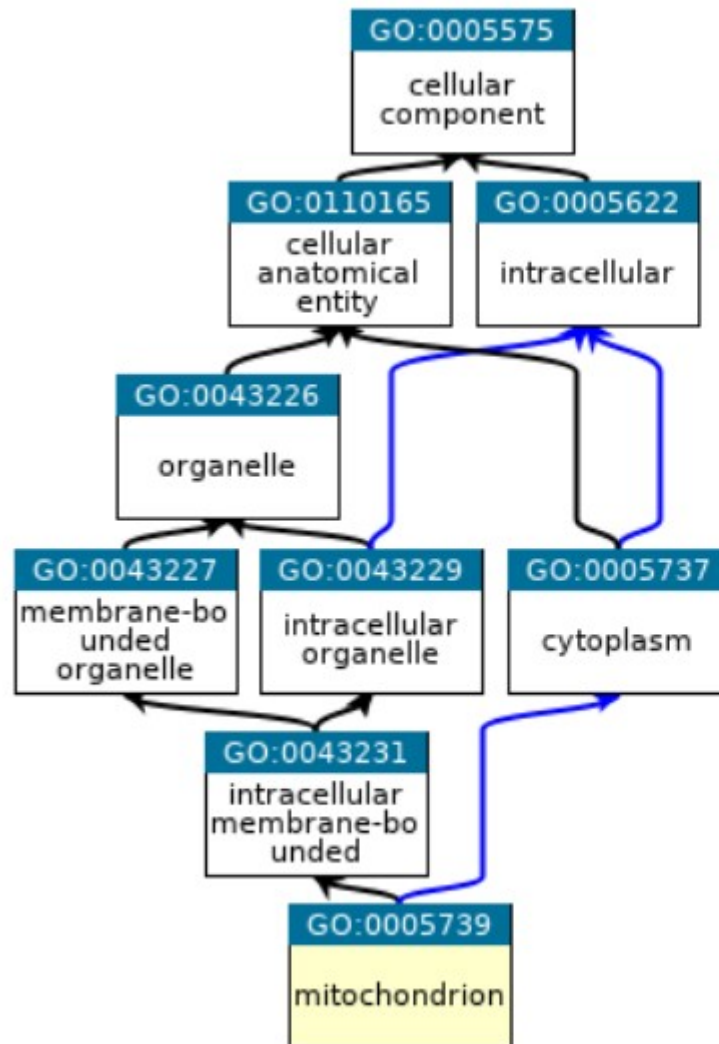
Gene ontology: GO



GENEONTOLOGY
Unifying Biology

- The Gene Ontology Consortium
 - <http://www.geneontology.org/>
- GO is a major bioinformatics initiative to **unify** the representation of gene and gene product attributes across all species.
- GO is the world's largest source of information on the functions of genes.
- It has a hierarchical structure
- Unified terminology
- 3 main parts:
 - *Molecular function* (i.e. RNA binding)
 - *Biological process* (i.e. reproduction)
 - *Cellular component* (i.e. mitochondria)

GO ancestor chart



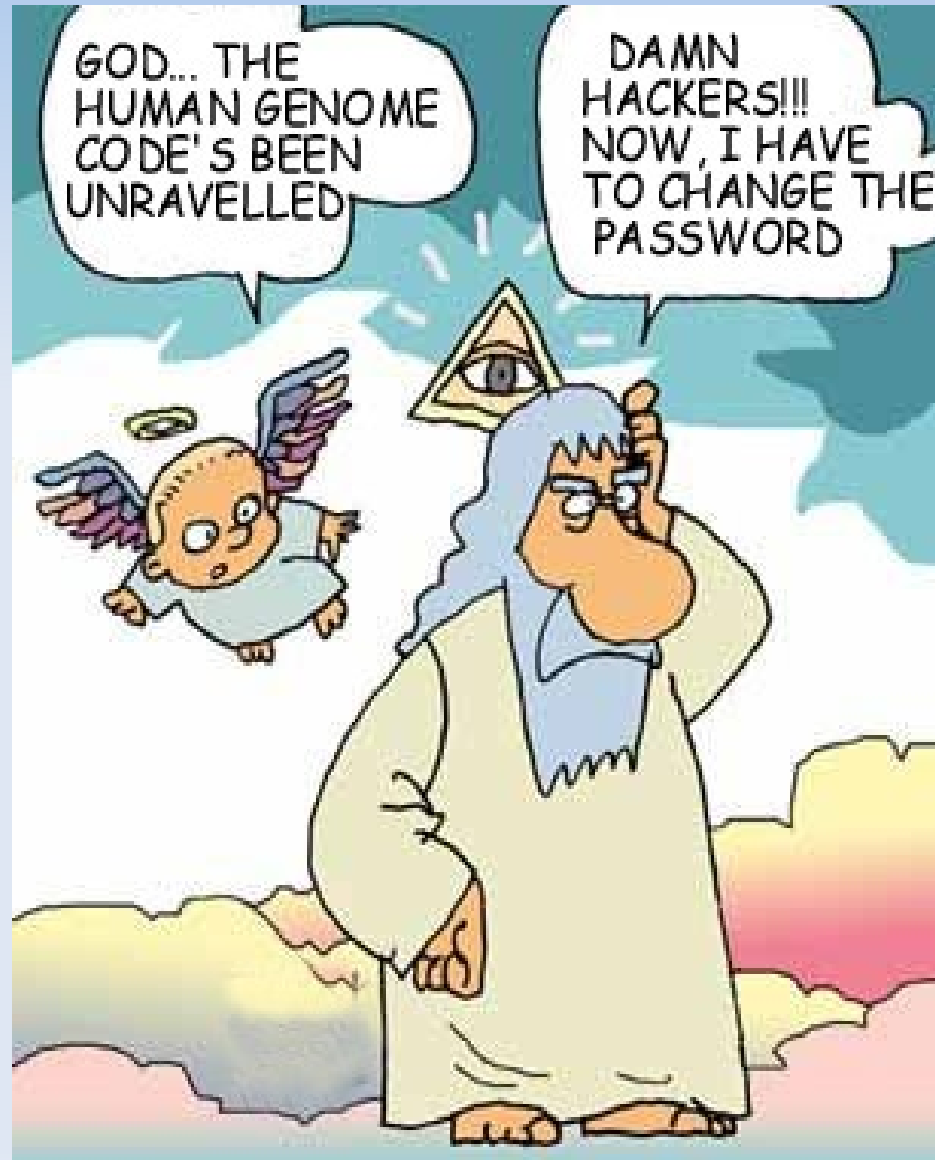
Expression Databases

- NCBI, GEO - Gene Expression Omnibus
 - <http://www.ncbi.nlm.nih.gov/geo/>
- EBI, ArrayExpress
 - <https://www.ebi.ac.uk/arrayexpress/>
- Microarray, RNAseq
- Sample, experiment base storing

Taxonomy database

- NCBI Taxonomy
- <https://www.ncbi.nlm.nih.gov/taxonomy/>
- A curated classification and nomenclature for all of the organisms in the public sequence databases.
 - It contains 452,352 species, and all together with higher and lower taxa: 604,541 entry.
- This currently represents about 10% of the described species of life on the planet.

Thank you for your attention



Proteins

- RASK (KRAS)
- ERK1 (MAPK3)
- JAK1
- IGF1R
- GSK3B
- AXIN1
- SMAD2
- NOTCH1

