HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

CODEJAM
2nd Workshop on Computational Biology

National and Kapodistrian University of Athens
School of Science
Departments of Biology
MSc of "Bionformatics-Computational Biology"

CODEJAM
Day 1 **Biological databases**

Dr. Nikolaos PAPANDREOU, Dr. Zoi LITOU, Dr. Ourania KONSTANTI & Assoc. Prof. Vassiliki ICONOMIDOU *(Dept. of Biology, NKUA)*

9-12am Introduction to Biological databases and ontologies

2-5 pm Lab applications of biological databases and ontologies

CODEJAM
Day 1 **Biological databases – Overview**

The data used from bioinformaticians are stored in databases in different file formats. This day1 course will be a guide to those biological resources, including what kinds of information they provide and how to access the data using the websites.

Instructors: Zoi Litou, Nikolaos Papandreou, Ourania Konstanti

CODEJAM
Day 1 **Biological databases – Learning objectives**

At the end of the course, you are expected to:

- Evaluate what the databases are and what you can do with them
- Navigate through the databases, explore the wide range of information provided and understand where it comes from
- Identify where different data files can be used

Prerequisites: Internet connection

CODEJAM
Day 1 **Biological databases – Introduction**

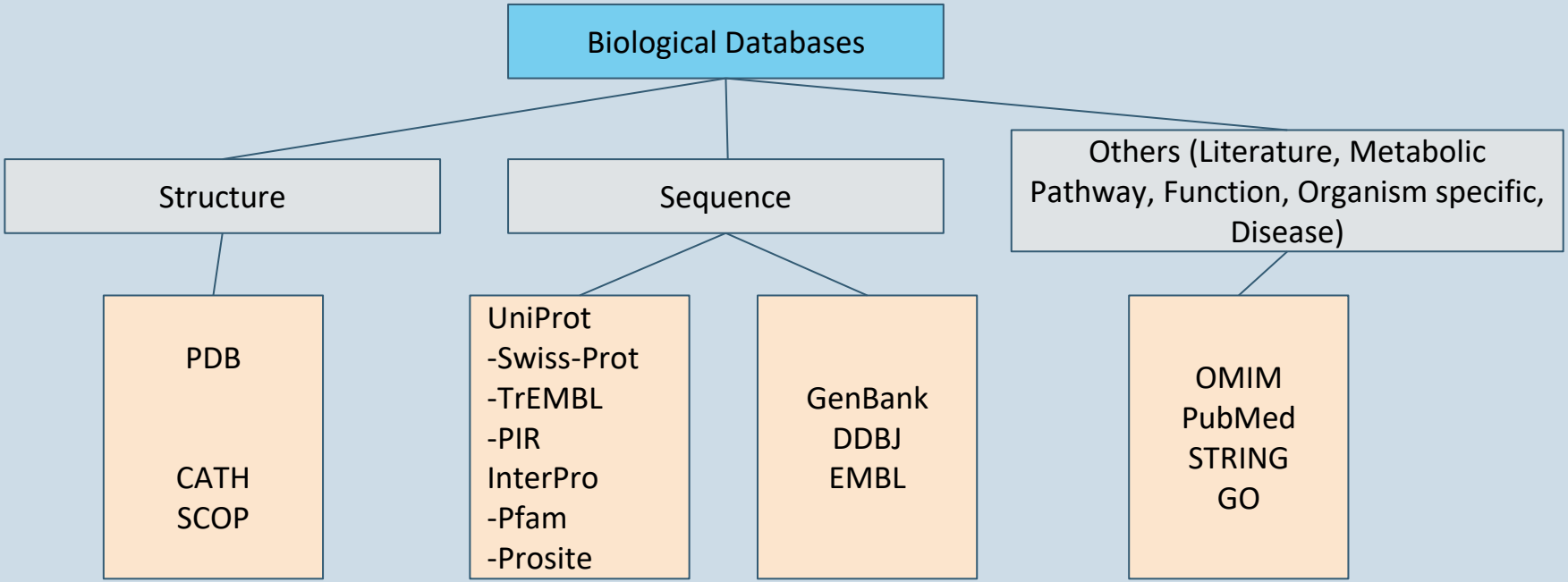
- Used to store and organize biological data to be retrieved easily
- Repositories with specific structure enabling the insertion and extraction of data
- The structure consist of files and tables
- Contains records and fields

CODEJAM
Day 1 **Biological databases – Types of Biological Databases (Source)**

Can be classified based on the information stored:

- Primary: experimentally derived data usually related to sequence or structure information (Protein-Nucleotide)
- Secondary: Contain information derived from primary databases
- Composite: Collections of several primary database resources, providing tools and software for data analysis

CODEJAM
Day 1 **Biological databases -Types of Biological Databases (Data)**



CODEJAM
Day 1 **Biological databases – Primary Nucleotide Databases**

Raw **nucleic acid sequence data** produced and submitted **by researches.**

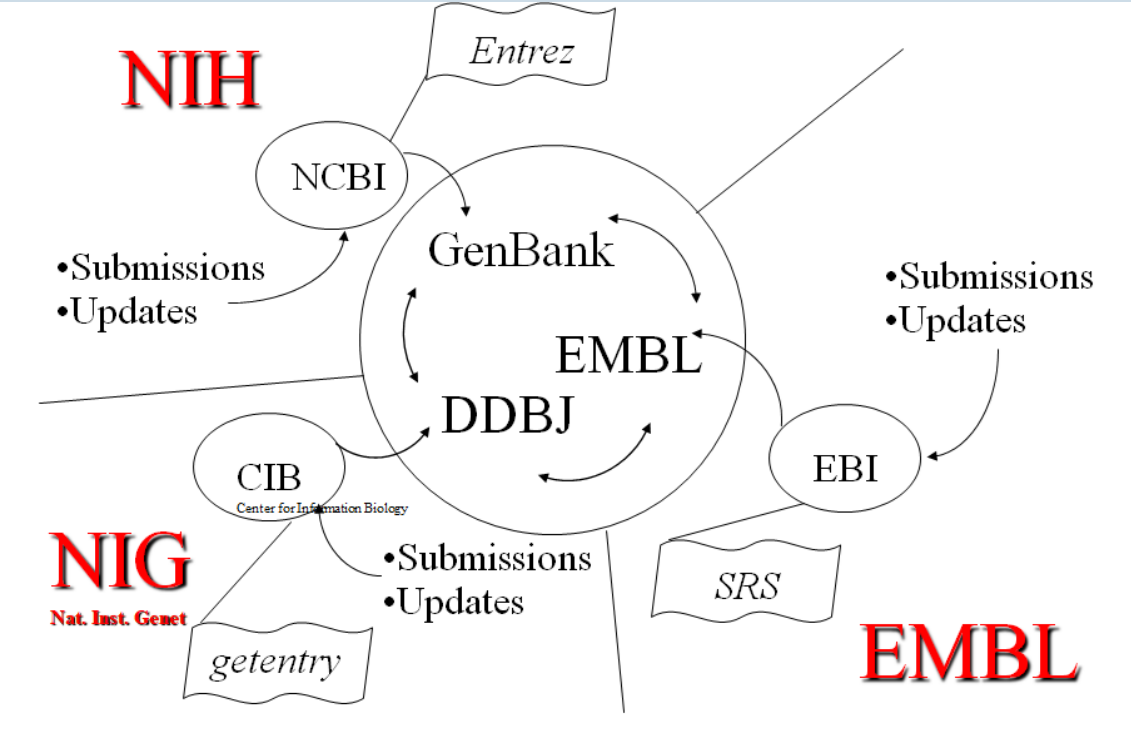
International Nucleotide Sequence Database Collaboration (INSDC), which is a joint effort between three primary databases: **GenBank, DDBJ,** and **EMBL**. These organizations work collaboratively to share sequence data from around the world on a daily basis and ensure that the data in each database is up-to-date and accurate.

GenBank

DDBJ

EMBL

CODEJAM
Day 1 **Biological databases – INSDC**



CODEJAM
Day 1 **Biological databases – GenBank**

- GenBank is a primary biological database managed by the National Center for Biotechnology Information (**NCBI**).
- It is an annotated collection of publicly available sequences, which includes information about genes, genomes, proteins, and other genetic elements.
- The GenBank flat file format is used to represent the sequence data and annotations in the database.
- GenBank accepts mRNA or genomic sequence data submitted by researches

CODEJAM
Day 1 **Biological databases – EMBL**

ENA
European Nucleotide Archive

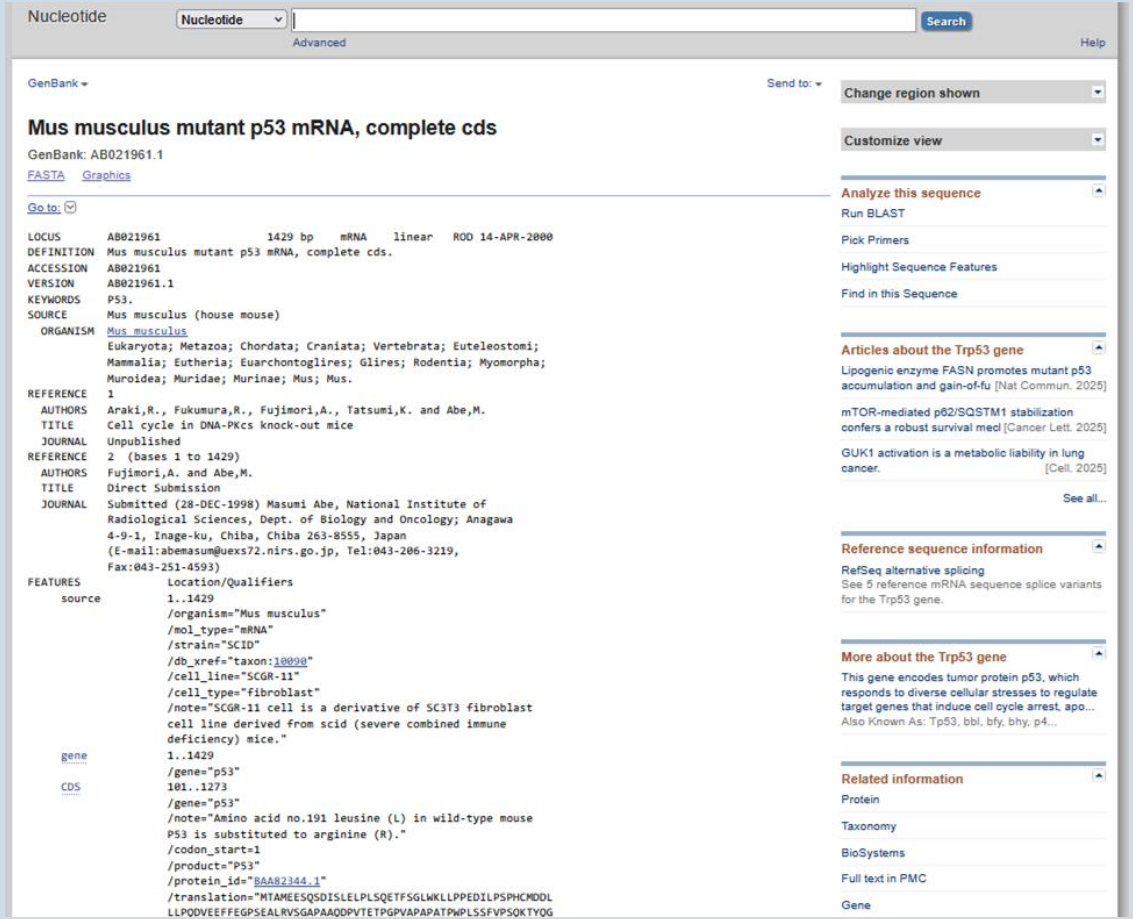
EMBL (European Molecular Biology Laboratory) is a collection of nucleotide sequence data that is maintained by the European Bioinformatics Institute (EBI). It is also a part of INSDC along with the GenBank and DDBJ databases.

EMBL's main focus is on the storage and distribution of nucleotide and protein sequences, as well as providing tools and resources for researchers to analyze and interpret this data

CODEJAM
Day 1 **Biological databases – DDBJ**

DDBJ (DNA Data Bank of Japan) is a primary database that collects and stores genetic information, mainly from Japanese researchers. They also receive and assign accession numbers to researchers from other countries

CODEJAM
Day 1 **Biological databases**
GenBank Flat File



Nucleotide | Nucleotide | Search | Advanced | Help

GenBank | Send to: | Change region shown | Customize view

Mus musculus mutant p53 mRNA, complete cds

GenBank: AB021961.1

FASTA Graphics

Go to:

```
LOCUS       AB021961               1429 bp    mRNA    linear   ROD 14-APR-2000
DEFINITION  Mus musculus mutant p53 mRNA, complete cds.
ACCESSION   AB021961
VERSION     AB021961.1
KEYWORDS    P53.
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
            Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1
  AUTHORS   Araki,R., Fukumura,R., Fujimori,A., Tatsumi,K. and Abe,M.
  TITLE     Cell cycle in DNA-PKcs knock-out mice
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1429)
  AUTHORS   Fujimori,A. and Abe,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-DEC-1998) Masumi Abe, National Institute of
            Radiological Sciences, Dept. of Biology and Oncology; Anagawa
            4-9-1, Inage-ku, Chiba, Chiba 263-8555, Japan
            (E-mail:abemasum@uexs72.nirs.go.jp, Tel:043-206-3219,
            Fax:043-251-4593)
FEATURES             Location/Qualifiers
     source          1..1429
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="SCID"
                     /db_xref="taxon:10090"
                     /cell_line="SCGR-11"
                     /cell_type="fibroblast"
                     /note="SCGR-11 cell is a derivative of SC3T3 fibroblast
                     cell line derived from scid (severe combined immune
                     deficiency) mice."
     gene            1..1429
                     /gene="p53"
     CDS             101..1273
                     /gene="p53"
                     /note="Amino acid no.191 leusine (L) in wild-type mouse
                     P53 is substituted to arginine (R)."
                     /codon_start=1
                     /product="P53"
                     /protein_id="BAA82344.1"
                     /translation="MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDL
                     LLPQDVEEFFEGPSEALRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQG
```

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Articles about the Trp53 gene
Lipogenic enzyme FASN promotes mutant p53 accumulation and gain-of-fu [Nat Commun. 2025]
mTOR-mediated p62/SQSTM1 stabilization confers a robust survival mecl [Cancer Lett. 2025]
GUK1 activation is a metabolic liability in lung cancer. [Cell. 2025]
See all...

Reference sequence information
RefSeq alternative splicing
See 5 reference mRNA sequence splice variants for the Trp53 gene.

More about the Trp53 gene
This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apo...
Also Known As: Tp53, bbl, bfy, bhy, p4...

Related information
Protein
Taxonomy
BioSystems
Full text in PMC
Gene

CODEJAM
Day 1 **Biological databases - GenBank Record**

- A column of keywords and a column of values (`ORGANISM` *`Saccharomyces cerevisiae`*).

•Basic info: gene name, accession number, taxonomy, references to literature.

•List of features found in the sequence

–The sequence itself is at the bottom, labeled ORIGIN

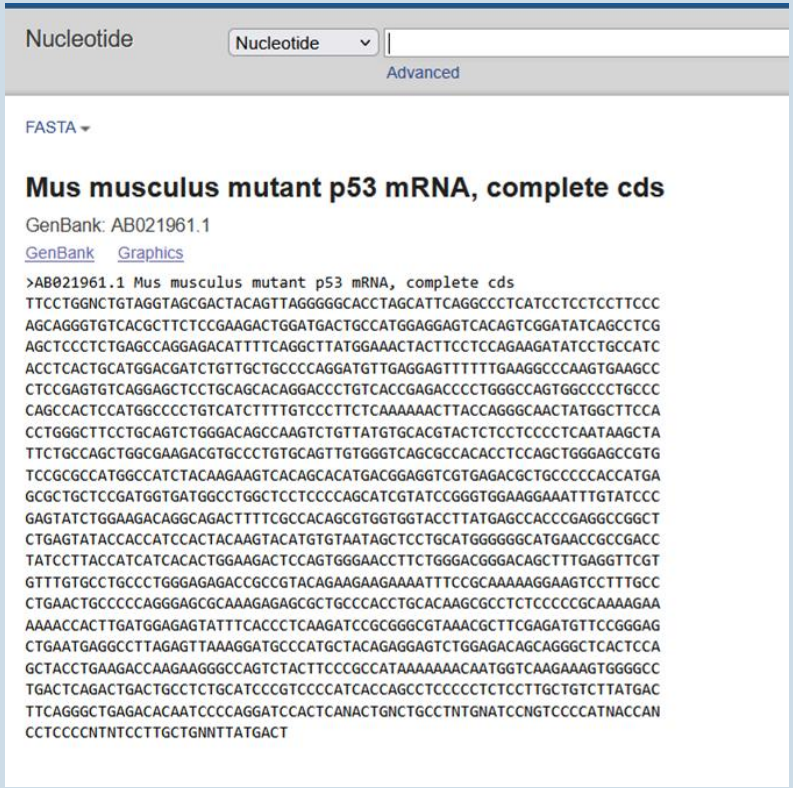
–The numbers refer to this sequence, sometimes with gaps and/or reverse-complement (feature on the opposite strand)

–"Gene" refers to the entire transcribed sequence

–"CDS" is just the protein-coding portion of the sequence

–"misc_feature" can be a wide variety of things

CODEJAM
Day 1 **Biological databases – FASTA Format**



CODEJAM
Day 1 **Biological databases – Development of protein databases**

Atlas of protein sequence and structure – Margaret Dayhoff (1966) first sequence database (pre-bioinformatics). Currently known as Protein Information Resource (PIR)

Protein data bank (PDB) – structural database (1972) remains most widely used database of structures

UniProt – The United Protein Databases (UniProt, 2003) is a central database of protein sequence and function created by joining the forces of the SWISS-PROT, TrEMBL and PIR protein database activities

CODEJAM
Day 1 **Biological databases – Primary Protein sequence Databases**

SWISS-PROT

-Manually curated

-high-quality annotations, less data

GenPept/TrEMBL

-Translated coding sequences from GenBank/EMBL

-Few annotations, more up to date

PIR (Protein Information Resources)

-Phylogenetic-based annotations

All 3 now combining efforts to form **UniProt** (http://www.uniprot.org)

CODEJAM
Day 1 **Biological databases – Swiss-Prot**

- Swiss-Prot created in 1986 from Amos Bairoch in SIB and then from Rolf Apweiler in EBI
- It has well annotated protein sequence data (function, keywords, description, classification)
- As non-redundant as possible
- Cross-references to other databases

CODEJAM
Day 1 **Biological databases**
Flat File

```
ID   CYS3_YEAST               STANDARD;      PRT;   393 AA.
AC   P31373;
DT   01-JUL-1993 (REL. 26, CREATED)
DE   CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).
GN   CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35.
OS   TAXONOMY
OC   SACCHAROMYCETACEAE; SACCHAROMYCES.
RX   CITATION
CC   -!- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +
CC            NH(3) + 2-OXOBUTANOATE.
CC   -!- COFACTOR: PYRIDOXAL PHOSPHATE.
CC   -!- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING
CC            L-CYSTEINE FROM L-METHIONINE.
CC   -!- SUBUNIT: HOMOTETRAMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES FAMILY.
CC   ---------------------------------------------------------------------
CC   DISCLAMOR
CC   ---------------------------------------------------------------------
DR   DATABASE cross-reference
KW   CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.
FT   INIT_MET      0          0
FT   BINDING     203          203            PYRIDOXAL PHOSPHATE (BY SIMILARITY).
SQ   SEQUENCE   393 AA;  42411 MW;  55BA2771 CRC32;
             TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPNRENL
             ERAVAALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE
             TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT
             FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP
             FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPGLKTHP NYDVVLKQHR
             DALGGGMISF RIKGGAEAAS KFASSTRLFT LAESLGGIES LLEVPAVMTH GGIPKEAREA
             SGVFDDLVRI SVGIEDTDDL LEDIKQALKQ ATN
//
```

CODEJAM
Day 1 **Biological databases – TrEMBL**

EMBL

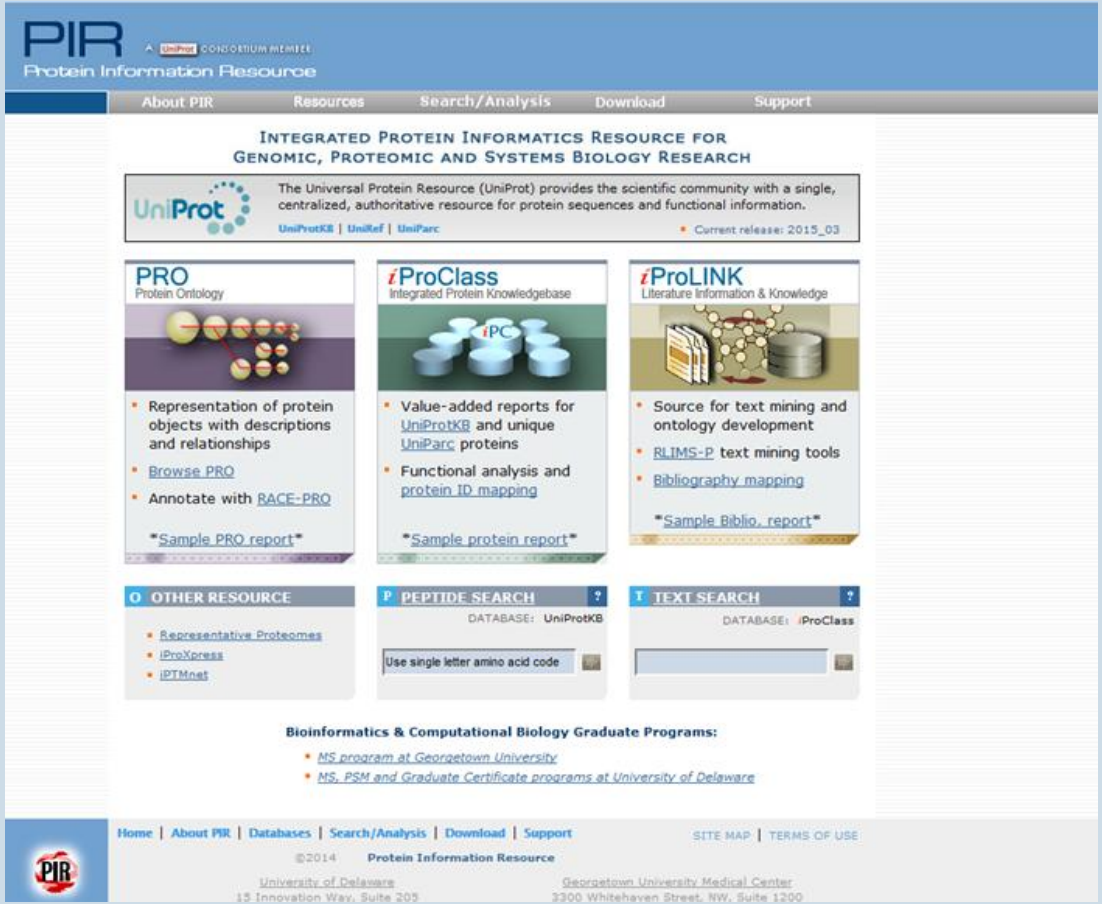
TrEMBL

Computer-annotated record derived

from a coding sequence (CDS)

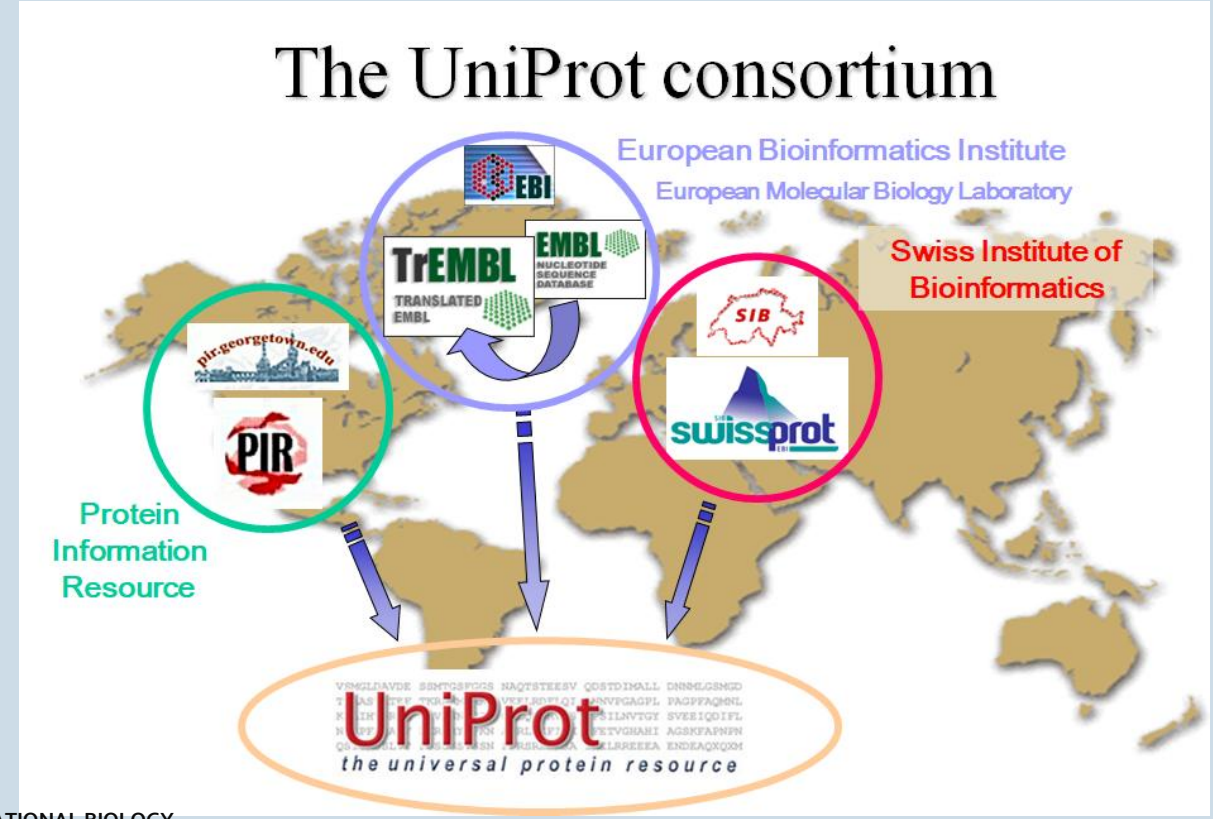
CODEJAM
Day 1 **Biological databases**
PIR

National Biomedical Research Foundation (NBRF)



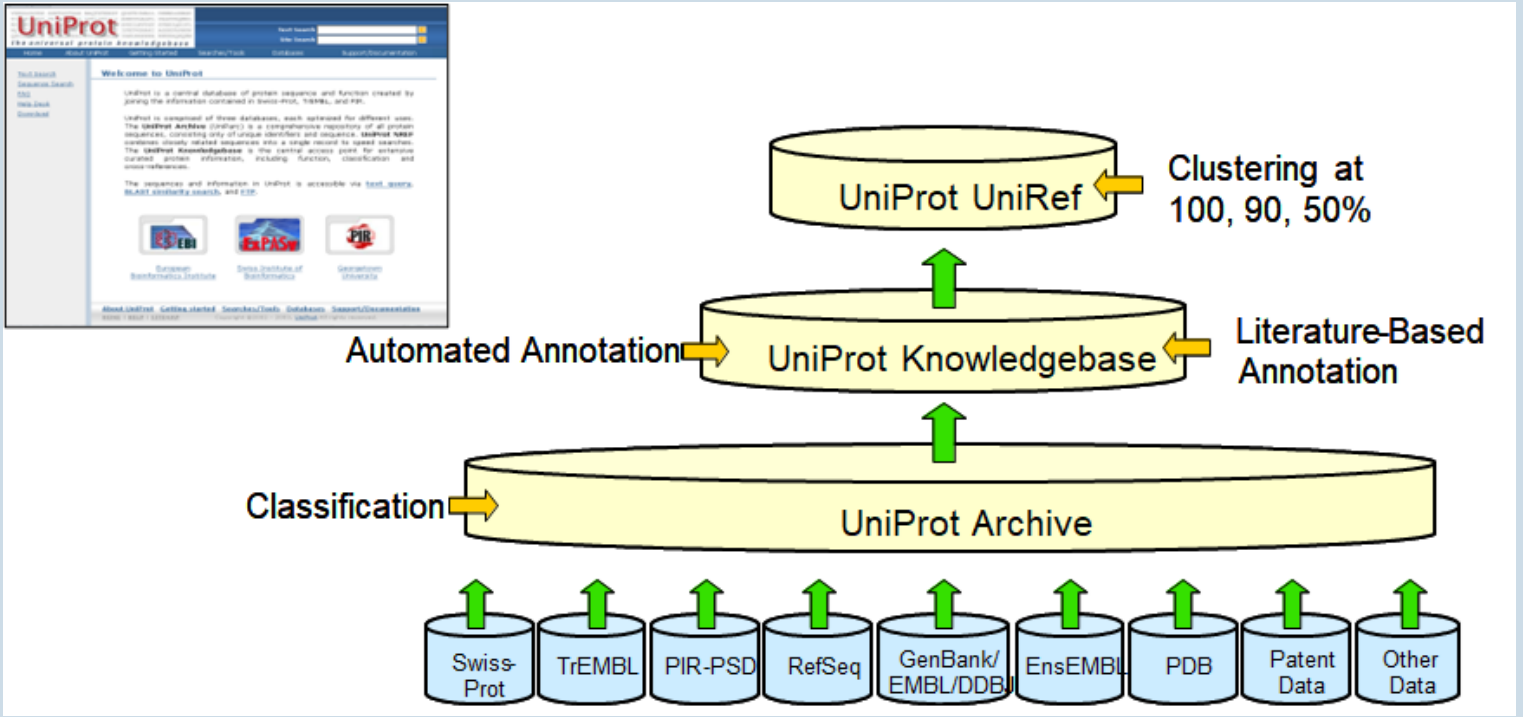
CODEJAM

Day 1 **Biological databases – Uniprot 2002**

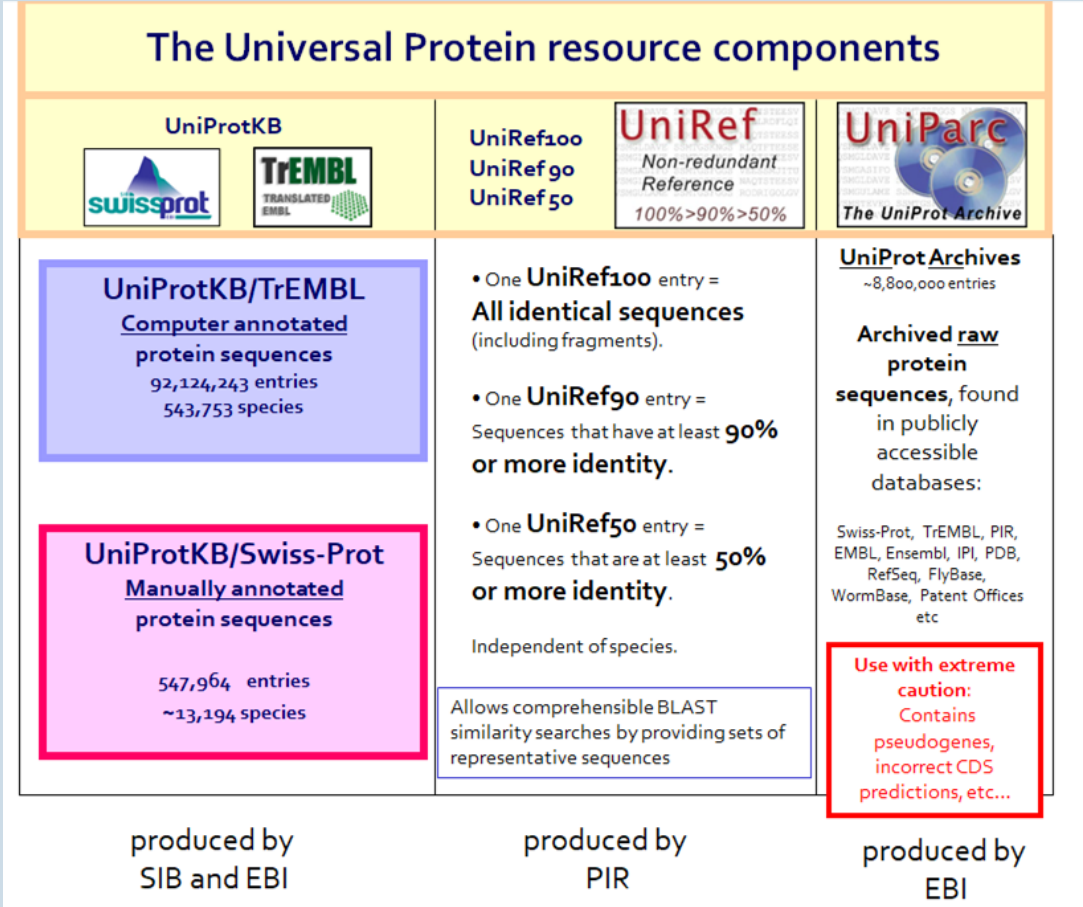


CODEJAM

Day 1 **Biological databases – Uniprot (Universal Protein Resource)**



CODEJAM
Day 1 **Biological databases**
Uniprot



CODEJAM

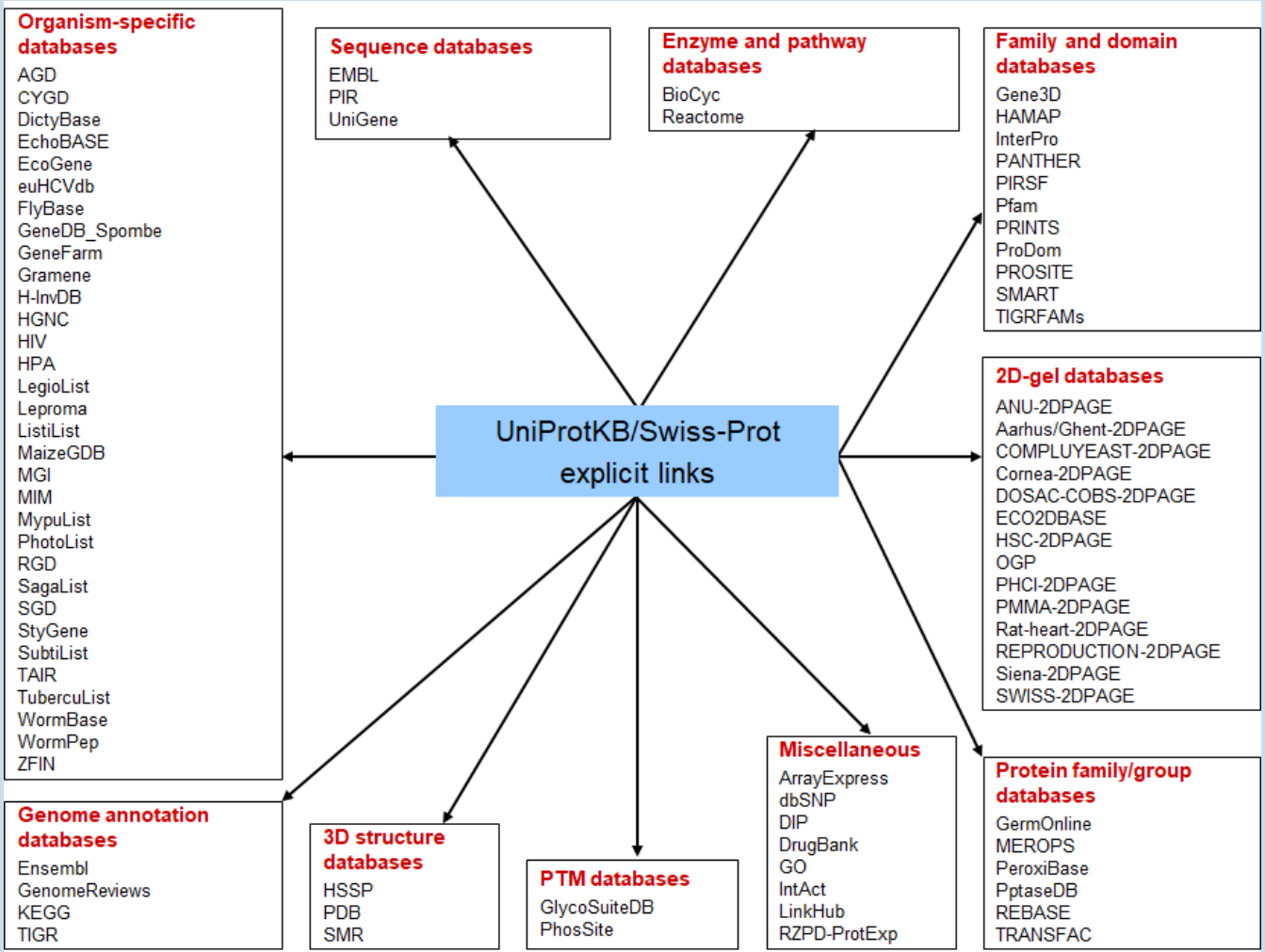
Day 1 **Biological databases – Uniprot**

No redundancy

Protein existence evidence

Annotation

Cross-reference


Organism-specific databases
AGD
CYGD
DictyBase
EchoBASE
EcoGene
euHCVdb
FlyBase
GeneDB_Spombe
GeneFarm
Gramene
H-InvDB
HGNC
HIV
HPA
LegioList
Leproma
ListiList
MaizeGDB
MGI
MIM
MypuList
PhotoList
RGD
SagaList
SGD
StyGene
SubtiList
TAIR
TubercuList
WormBase
WormPep
ZFIN
Sequence databases
EMBL
PIR
UniGene
Enzyme and pathway databases
BioCyc
Reactome
Family and domain databases
Gene3D
HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs
UniProtKB/Swiss-Prot explicit links
2D-gel databases
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
Cornea-2DPAGE
DOSAC-COBS-2DPAGE
ECO2DBASE
HSC-2DPAGE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
REPRODUCTION-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
Genome annotation databases
Ensembl
GenomeReviews
KEGG
TIGR
3D structure databases
HSSP
PDB
SMR
PTM databases
GlycoSuiteDB
PhosSite
Miscellaneous
ArrayExpress
dbSNP
DIP
DrugBank
GO
IntAct
LinkHub
RZPD-ProtExp
Protein family/group databases
GermOnline
MEROPS
PeroxiBase
PptaseDB
REBASE
TRANSFAC

CODEJAM
Day 1 **Biological databases -Uniprot Summary**

- Swiss-Prot is the **non redundant**, **manually annotated** and **highly cross-referenced** section of the UniProt Knowledgebase
- Be aware of the differences between UniProtKB/TrEMBL and UniProtKB/Swiss-Prot
  - Computer vs. Human
  - Redundant vs. Non-redundant
- **Always** cite the Accession number, not the entry name
  - The AC is stable
  - The entry name might change (ID)

CODEJAM
Day 1 **Biological databases - Secondary protein Databases**

Can store:

Conserved domain

Active site residues

Motifs

Patterns-Profiles

Interpro

CODEJAM
Day 1 **Biological databases – Primary Protein Structure Databases**

PDB (Protein DataBank) -continue

CODEJAM
Day 1 **Biological databases – Primary Protein Structure Databases**

Citation of images used:

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/

The UniProt Consortium

UniProt: the Universal Protein Knowledgebase in 2025

Nucleic Acids Res. 53:D609–D617 (2025)