Lesson 7.
Content

1. Biological databases.

Availability of biological data

In the early '50 Frederick Sanger determined the first complete amino acid sequence (the two polypeptide chains of bovin insulin, A and B)
After 10 years the first attempts were made to create a database of protein sequences

Nowadays the number of known protein sequences is in the order of hundreds of millions, 10^8 (≅230 million in UniProt at October 2022)

The first protein 3D structures to be solved were hemoglobin by Max Perutz and myoglobin by John Kendrew, in 1958

Nowadays the number of known protein 3D structures in the Protein Data Bank (PDB) are ≅200,000 (October 2022)

Relevance and structure of biological databases

Biological data and therefore databases (DBs) which contain it constitute the backbone of bioinformatics research

There are many different ways to design databases, both in terms of the ways the information is stored and the ways it can be retrieved and analysed

Modern databases store data and contain complex technology to store it in a structured manner, to allow it to be efficiently INPUT, ACCESSED and ANALYSED

Databases require databases management system (DBMS), a software to control the DB

Relevance and structure of biological databases

Databases contain entries, made of **data** (e.g. the sequence in a sequence DB) and **annotation** (e.g. organism, gene location and name, name and sequence of the encoded protein, etc.)

```
ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX
AC   J00231;
XX
SV   J00231.1
XX
DT   13-JUN-1985 (Rel. 06, Created)
DT   02-JUL-1999 (Rel. 60, Last updated, Version 7)
XX
DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW   C-region; gamma heavy chain disease protein;
KW   gamma3 heavy chain disease protein; heavy chain disease; hinge exon;
KW   immunoglobulin gamma-chain; immunoglobulin heavy chain;
KW   secreted immunoglobulin; V-region.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheri
OC   Primates; Catarrhini; Hominidae; Homo.
```

```
FH   Key             Location/Qualifiers
FH
FT   source          1..1089
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
FT                   /map="14q32.33"
FT   mRNA            <1..1089
FT                   /note="gamma3 mRNA"
FT   gene            23..964
FT                   /gene="IGHG3"
FT   CDS             23..964
FT                   /codon_start=1
FT                   /db_xref="SWISS-PROT:P01860"
FT                   /note="OMM protein (Ig gamma3) heavy chain"
FT                   /gene="IGHG3"
FT                   /protein_id="AAA52805.1"
FT                   /translation="MKXLWFFLLLVAAPRWVLSQVHLQESGPGLGKPPELK
FT                   TCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPXCP
FT                   SVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPXVQFKNYVDGVEVHNAKT
FT                   STFRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPXXXXX
FT                   ENTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYNTTPPMLDSDGSFFLY
FT                   RMQQGNIFSCSVMHEALHNRYTQKSLSLSPGK"
```

Feature Table

ANNOTATION

```
SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
     cctggacctc ctgtgcaaga acatgaaaca nctgtggttc ttccttctcc tggtggcagc
60
     tcccagatgg gtcctgtccc aggtgcacct gcaggagtcg ggcccaggac tggggaagcc
120
     tccagagctc aaaaccccac ttggtgacac aactcacaca tgcccacggt gcccagagcc
180
     caaatcttgt gacacacctc ccccgtgccc acggtgccca gagcccaaat cttgtgacac
240
     acctccccca tgcccacggt gcccagagcc caaatcttgt gacacacctc ccccgtgccc
300
     nnngtgccca gcacctgaac tcttgggagg accgtcagtc ttcctcttcc ccccaaaacc
360
     caaggatacc cttatgattt cccggacccc tgaggtcacg tgcgtggtgg tggacgtgag
420
     ccacgaagac ccnnnngtcc agttcaagtg gtacgtggac ggcgtggagg tgcataatgc
480
     caagacaaag ctgcgggagg agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac
540
     cgtcctgcac caggactggc tgaacggcaa ggagtacaag tgcaaggtct ccaacaaagc
600
     cctcccagcc cccatcgaga aaaccatctc caaagccaaa ggacagcccn nnnnnnnnnn
660
     nnnnnnnnnn nnnnnnnnnn nnnnngagga gatgaccaag aaccaagtca gcctgacctg
720
     cctggtcaaa ggcttctacc ccagcgacat cgccgtggag tgggagagca atgggcagcc
780
     ggagaacaac tacaacacca cgcctcccat gctggactcc gacggctcct tcttcctcta
840
     cagcaagctc accgtggaca agagcaggtg gcagcagggg aacatcttct catgctccgt
```

DATA

Example of entry from a nucleotide sequence DB (GenBank)

The structure of databases: flat-file format

In general a database structure consists of files or tables each containing **records** and **fields**

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

fields

Example of a very simple database table: a single page with a contact list with 3 records and 3 fields per record

The structure of databases: flat-file format

In general a database structure consists of files or tables each containing **records** and **fields**

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

records

Example of a very simple database table: a single page with a contact list with 3 records and 3 fields per record

The structure of databases: flat-file format

In general a database structure consists of files or tables each containing **records** and **fields**

(B) GenBank Flat-File Format

```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and
             Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
           Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
```

fields

Example of a GenBank record in a flat-file format

The structure of databases: relational databases

In relational DBs, the most commonly used for biological information, data is stored within a number of tables

Each table consists of **records** and **fields** and is linked to other tables by a shared field called **key**, unique to each record

fields

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

records

fields

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

records

*This example of relational database consists of two tables with different fields, where records are connected by an identifier (**key**)*

The structure of databases: relational databases

In relational DBs, the most commonly used for biological information, data is stored within a number of tables

Each table consists of **records** and **fields** and is linked to other tables by a shared field called **key**, unique to each record

fields

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

records

keys

fields

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

records

*This example of relational database consists of two tables with different fields, where records are connected by an identifier (**key**)*

The structure of databases: relational databases

In relational DBs, a set of operators is provided that allows to manipulate and analyse the data (matematical, logical etc.)

Usually, results produced by application of operators on the data are displayed in new tables

fields

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

records

keys

fields

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

records

*This example of relational database consists of two tables with different fields, where records are connected by an identifier (**key**)*

Searching DBs

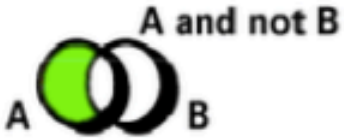
Boolean operators help to focalize DB searches:

Boolean Operators

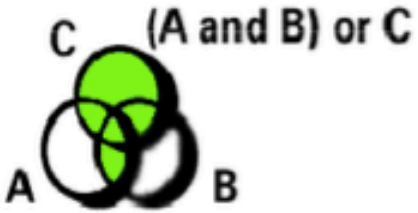
- AND (&)

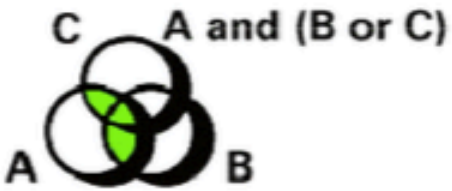
- OR (|)

- NOT(!)



Complex queries





Searching DBs

Boolean operators help to focalize DB searches:

e.g. "human" will list all the entries containing any information on homo sapiens

"human liver alcohol dehydrogenase" (human AND liver AND alcohol AND dehydrogenase) will list only entries specific for the enzyme

Databases usually provide **fields** which can be selected in order to focalize a search, e.g. author ([au]), organism [OS], etc

The structure of databases: SQL

In relational DBs management systems, operators are written in query-specific languages, such as the **Structured Query Language** (**SQL**)

Examples of SQL application on a relational database

example-1

Query 1

```
SELECT protein-code, protein-name
FROM protab1
WHERE species-origin = 'Bovine';
```

P1001	Hemoglobin

example-2

Query 2

```
SELECT protab1.protein-name, protab2.protein-sequence
FROM protab1, protab2
WHERE protab1.protein-code = protab2.protein-code
AND protab1.protein-code = 'P1002';
```

Hemoglobin	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…

The structure of databases: XML

eXstensible Markup Language (**XML**) is a powerful system for marking up (annotating) data

It is one of many markup languages, including HTML (hypertext markup language), commonly used to write web pages

The hallmark of these languages is the use of identifiers called tabs, which can enclose sections of data, e.g. <language>XML</language>

XML has mechanisms allowing arbitrary tags to be used, thus having the flexibility to define bespoke (tailored) data classifications

XML uses plain file format, which makes it portable and accessible

Many bioinfomatics DBs are being made available in XML format, although a master copy is often maintained e.g. as relational DBs

Primary and secondary (derived) biological DBs

PRIMARY DBs: contain primary data, i.e. experimental data & annotations

e.g. nucleotide sequence coding, non coding

SECONDARY DBs: contain secondary data, i.e. analyses of general interest derived from primary DBs & annotations

e.g. collections of conserved sequence motifs functional, non functional

A number of centers have been funded to provide access to a large number of major databases in an integrated environment

More on DNA sequences in DBs

There are different types of DNA sequences

Genomic (chromosomal) sequences:

from genome sequencing projects – in GenBank

include noncoding regions, introns, control regions etc.

cDNA (complementary DNA) sequences:

obtained from mRNA sequences by reverse transcription (copying mRNA in DNA)

represent genes actually expressed in a specific cell/tissue at a given condition

do not include gene regions not transcribed in mRNA (introns, control regions etc.)

ESTs (expressed sequence tags):

obtained from mRNA sequences by reverse transcription (copying mRNA in DNA)

are partial cDNA sequences, representing expressed genes – useful for genomes scanning

Examples of primary biological databases

..aatgcatgccaatg
ccatccgcatcgat..

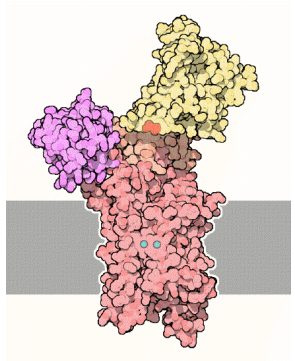
EMBL Nucleotide Sequence Database

GenBank

..QEDARTSCG
AILNQRYWI..

UniProt/Swiss-Prot

Protein Information Resource (**PIR**)



RCSB PDB
PROTEIN DATA BANK

ensembl

PubMed

Examples of primary biological databases: GenBank

NCBI HomePage - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/

NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search Gene for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Influenza Virus Resource

The Influenza Virus Resource enables comparison of influenza virus strains and provides a reference for viral sequences. The resource contains data from the NIAID Influenza Genome Sequencing Project and GenBank, as well as pre-computed alignments of flu sequences.

Entrez Gene

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources

Examples of primary biological databases: UniProt

Text Search UniProt Knowledgebase - UniProt [the Universal Protein Resource] - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ebi.uniprot.org/uniprot-srv/index.do

UniProt
the universal protein knowledgebase

Home > Database > UniProt Knowledgebase Search

hosted by

Text Search UniProt Knowledgebase

Home | About UniProt | Getting Started | Searches/Tools | Databases | Support/Documentation

- Text Search
- Power Search
- Warehouse
  - Prediction Search
  - InterPro Search
  - CluSTr Search
  - Entry List Search
- Data Set Manager
- BLAST
- FAQ
- Help Desk
- Download

Text Search UniProt Knowledgebase

Text Search | Power Search | Warehouse | UniProt | UniParc | UniRef

Insert a query

1929823 Entries In UniProt Release 5.1 | Help On Your Query | Search & View | Reset

A common question: **How can I build queries with logical operators?**

Examples of primary biological databases: UniProt

UniProt constists of two parts:

October '22

1 entry per protein

Non redundant, high-quality manual annotation - reviewed

1 entry per nucleotide submission

Redundant, automatically annotated – unreviewed

UniProt

```
ID   Q9XSK1      PRELIMINARY;      PRT;   142 AA.
AC   Q9XSK1;
DT   01-NOV-1999 (TrEMBLrel. 12, Created)
DT   01-NOV-1999 (TrEMBLrel. 12, Last sequence update)
DT   01-NOV-1999 (TrEMBLrel. 12, Last annotation update)
DE   II ALPHA4 HAEMOGLOBIN CHAIN.
GN   II ALPHA4.
OS   Bubalus bubalis (Water buffalo).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
OC   Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovoidea; Bovidae;
OC   Bovinae; Bubalus.
RN   [1]
RP   SEQUENCE FROM N.A.
RA   Ferranti P., Rullo R., Zappacosta F., Vincenti D., Masala B.,
RA   Di Luccia A.;
RT   "River buffalo (Bubalus bubalis L.) alpha globin gene and chain
RT   sequences: evolutionary structural relationship among some ruminant
RT   species.";
RL   Submitted (MAY-1999) to the EMBL/GenBank/DDBJ databases.
DR   EMBL; AJ242734; CAB43765.1; -.
SQ   SEQUENCE   142 AA;  15171 MW;  99F24F6E285C758F CRC64;
     MVLSAADKSN VKAAWGKVGG HAADYGAEAL ERMFLSFPTT KTYFPHFDLS HGSAQVKGHG
     AKVANALTKA VGHLDDLPGA LSELSDLHAH KLRVDPVNFK LLSHSLLVTL ASHLPNDFTP
     AVHASLDKFL ASVSTVLTSK YR
//
```

Sequence in FASTA format

>Q9XSK1
MVLSAADKSNVKAAWGKVGGHAADYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
AKVANALTKAVGHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPNDFTP
AVHASLDKFLASVSTVLTSKYR

UniProt

For a protein sequence

Annotation may also report:

regions playing a function, interacting with other molecules, featuring a certain secondary structure, involved in disulfide bridges, *etc. etc.*

```
ID   Q9XSK1      PRELIMINARY;      PRT;   142 AA.
AC   Q9XSK1;
DT   01-NOV-1999 (TrEMBLrel. 12, Created)
DT   01-NOV-1999 (TrEMBLrel. 12, Last sequence update)
DT   01-NOV-1999 (TrEMBLrel. 12, Last annotation update)
DE   II ALPHA4 HAEMOGLOBIN CHAIN.
GN   II ALPHA4.
OS   Bubalus bubalis (Water buffalo).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
OC   Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovoidea; Bovic
OC   Bovinae; Bubalus.
RN   [1]
RP   SEQUENCE FROM N.A.
RA   Ferranti P., Rullo R., Zappacosta F., Vincenti D., Masala B.,
RA   Di Luccia A.;
RT   "River buffalo (Bubalus bubalis L.) alpha globin gene and chai
RT   sequences: evolutionary structural relationship among some rur
RT   species.";
RL   Submitted (MAY-1999) to the EMBL/GenBank/DDBJ databases.
DR   EMBL; AJ242734; CAB43765.1; -.
SQ   SEQUENCE   142 AA;  15171 MW;  99F24F6E285C758F CRC64;
     MVLSAADKSN VKAAWGKVGG HAADYGAEAL ERMFLSFPTT KTYFPHFDLS HGSAQVH
     AKVANALTKA VGHLDDLPGA LSELSDLHAH KLRVDPVNFK LLSHSLLVTL ASHLPNI
     AVHASLDKFL ASVSTVLTSK YR
//
```

Sequence in FASTA format

```
>Q9XSK1
MVLSAADKSNVKAAWGKVGGHAADYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
AKVANALTKAVGHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPNDFTP
AVHASLDKFLASVSTVLTSKYR
```

Example of primary database

http://www.rcsb.org/pdb/Welcome.do;jsessionid=O9zmva+Ao1jjCPYGhF3Ikg**

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Mar 07, 2006 there are 35460 Structures | PDB Statistics

Contact Us | Help | Print Page

PDB ID or keyword | Author | SEARCH | Advanced Search

Home | Search

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Tissue Factor

Blood performs many essential jobs in your body: it transports oxygen and nutrients, it protects your cells from infection, and it carries hormones and other messages from place to place in your body. But since blood is a liquid that is pumped under pressure, we must protect

NEWS

- Complete News
- Newsletter
- Discussion Forum

07-Mar-2006
RCSB PDB Focus: Frequently Asked Questions

The Frequently Asked Questions page answers a number of common queries about the new RCSB PDB site

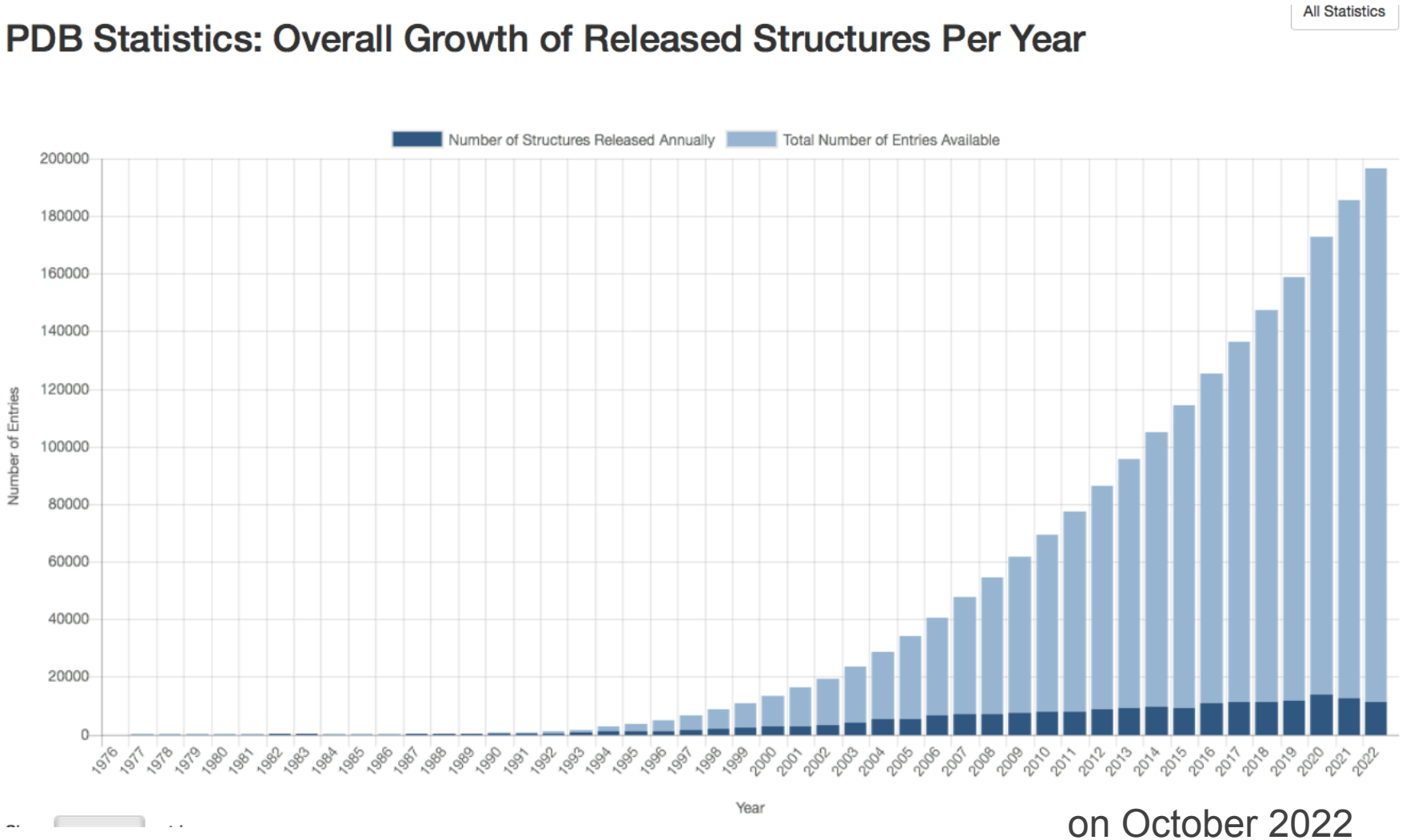
- Full Story ...

28-Feb-2006
RCSB PDB Exhibit News

21-Feb-2006
Virtual Reality Environment Highlights PDB Structures

Some statistics on the Protein Data Bank (PDB)

The Protein Data Bank has been established in 1971. Since then, the number of available 3D structures has grown exponentially



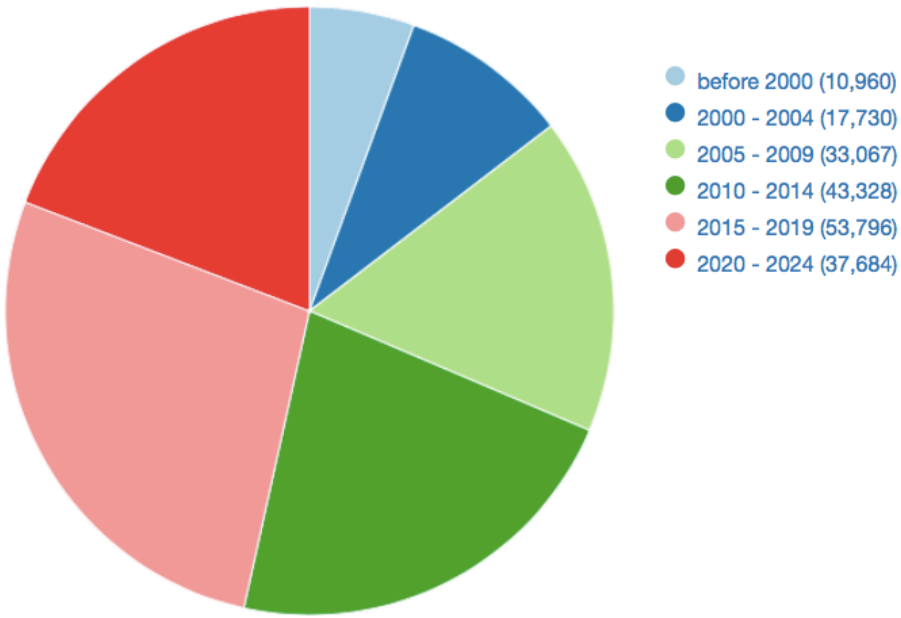
on October 2022

Some statistics on the Protein Data Bank (PDB)

The Protein Data Bank has been established in 1971

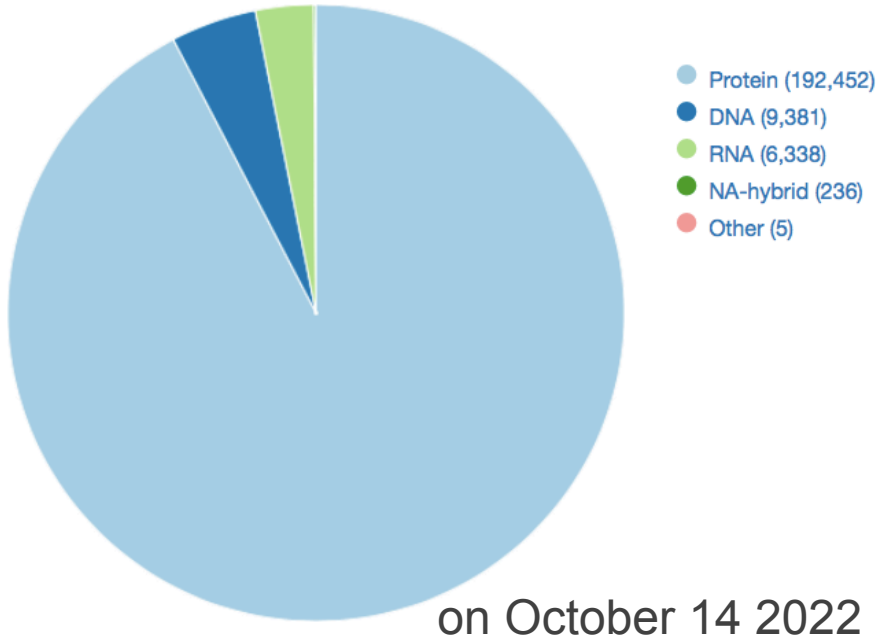
By Release Date

The year a structure entry was released in the PDB archive.



By Polymer Entity Type

PDB structures may contain multiple entities of different macromolecular types.



on October 14 2022

The 3D structures released in the last 3 years (2020-2022) are roughly 4 times those released in the first 30 years (1971-2000)

The 3D structures in the PDB are mostly for proteins but also for DNA and RNA molecules

Examples of primary biological databases: Ensembl

ensembl

Universal source of information on the human genome and other eukaryote genomes (over 200 species to date)

DATA: genomes: genes, SNPs (Single Nucleotide Polymorphisms), repetitive sequences

ANNOTATIONS: coding regions, for each gene proofs supporting its identification

CONNECTIONS: to other DBs and software

Examples of secondary (derived) biological databases

Name	*Derived from*	*Contains*
EMEST	EST Database	Collection and alignment of EST seqs
DSSP	Protein Data Bank (PDB)	Protein secondary structure assignments
HSSP	Protein Data Bank (PDB) & protein sequences DBs	Alignments of protein seqs of known structure with all similar seqs
FSSP, SCOP, CATH	Protein Data Bank (PDB)	Structural classification
3Dee	Protein Data Bank (PDB)	Protein domains definition
Pfam, Prints, BLOCKS	Protein sequences DBs	Alignments of homologous protein families & domains
Prodom	Protein sequences DBs	Alignments of similar protein domains
PROSITE	Protein sequences DBs	Sequence patterns
OMIM	Genomic databases	Genes & associated genetic diseases
LocusLink	Genomic databases	Genetic loci

Protein domains

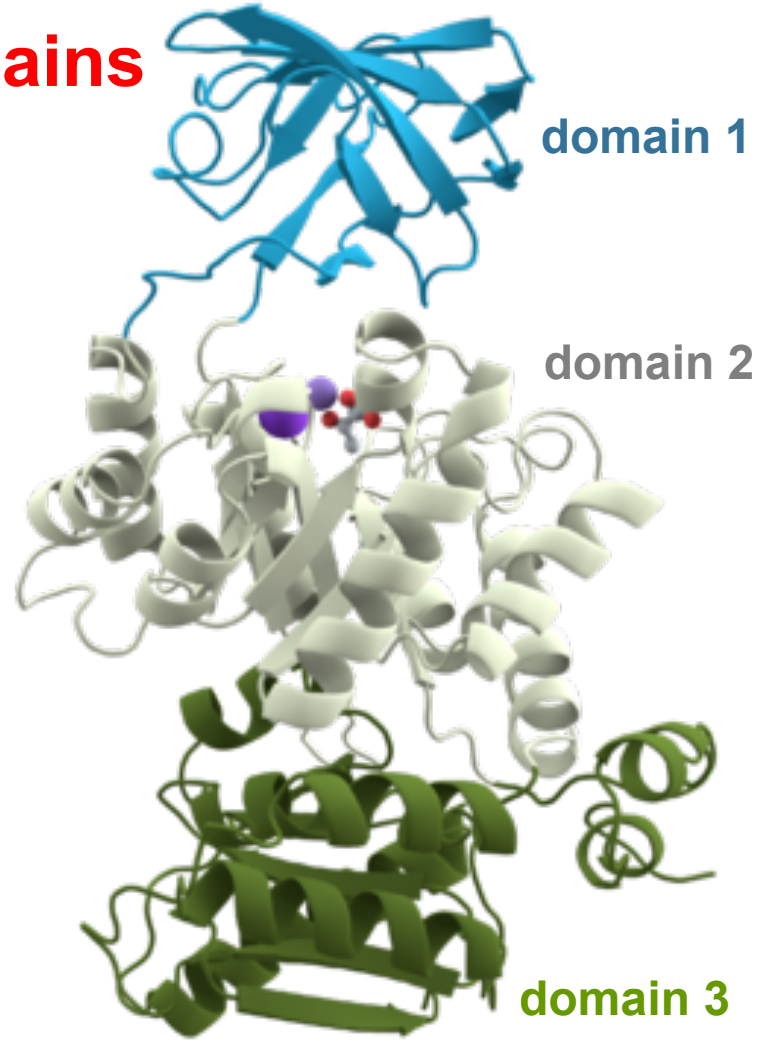
Many proteins consist of several **domains**, which usually form **functional units**

A protein domain is a region of a protein chain that is self-stabilizing and that **folds independently from the rest**

A protein domain is usually ≈50 to ≈250 amino acid long

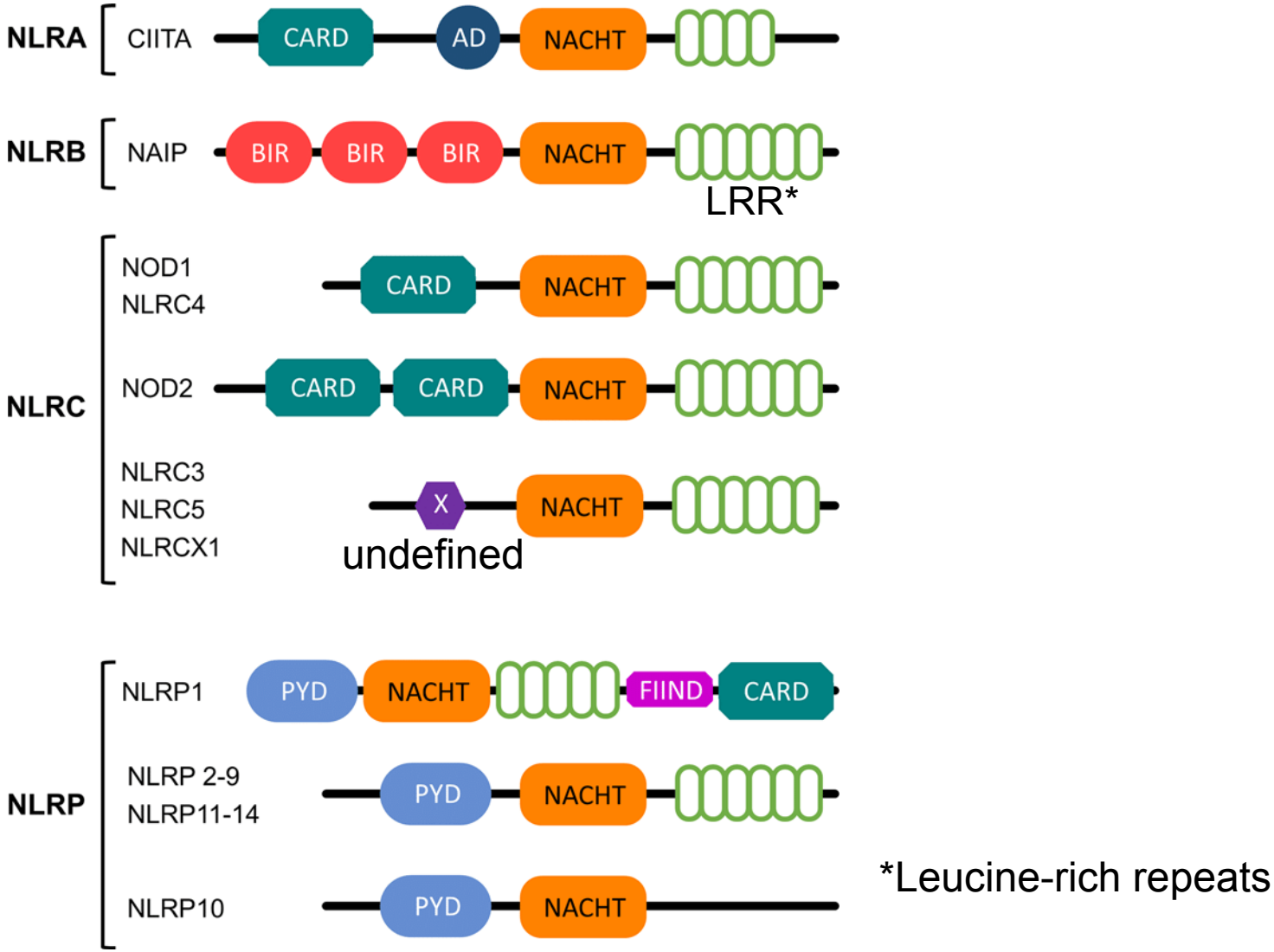
A domain may appear in a variety of different proteins

Molecular evolution uses domains as building blocks in different proteins



***Pyruvate kinase**: contains an all-β nucleotide binding domain (blue), an α/β-substrate binding domain (grey) and an α/β-regulatory domain (green), connected by several linkers.*

Protein domains as building blocks for different proteins



*Example of domains combination in the **N**OD-**L**ike **R**eceptors proteins*

Examples of secondary (derived) biological databases

Pfam: Pfam Home Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.sanger.ac.uk/Software/Pfam/

wellcome trust sanger institute

Pfam

Search Pfam

RSS | Pfam Home | Search by | Browse by | FTP | iPfam | Help | About

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 18.0 and 19.0, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.

75% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 7973 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

Version 19.0

December 2005, **8183** families

Sequence coverage Pfam-A : 75%
Sequence coverage Pfam-B : 19%
Other

Web feed

You can use the RSS feed to keep updated about Pfam releases

XML RSS

Enter your keyword(s) here

Go Example

Enter a SWISS-PROT 48.1 or TrEMBL 31.1 name or accession number

Go Example

Search The Web

Find a Web page containi

Search

Brought to you by MSN Search

Search for other items: Files or Folders Computer People

© 2006 Microso MSN Privacy

Pfam is a collection of protein families and domains
Contains multiple alignments and HMM profiles

Examples of secondary (derived) biological databases

Pfam: Pfam Home Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.sanger.ac.uk/Software/Pfam/

wellcome trust sanger institute

Pfam

Search Pfam

RSS | Pfam Home | Search by | Browse by | FTP | iPfam | Help | About

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or o .0, click here.

Pfam can be used to view the domain organisation of that a single protein can belong to several Pfam families.

IMPDH CBS IMPDH

75% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 7973 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam with

Version 19.0

December 2005, **8183** families

B A

Sequence coverage Pfam-A : 75%
Sequence coverage Pfam-B : 19%
Other

Web feed

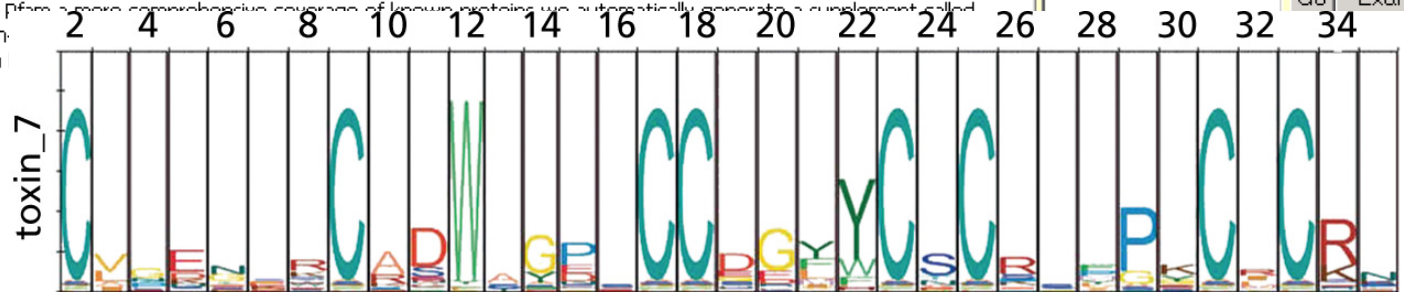
You can use the RSS feed to keep updated about Pfam releases

XML RSS

Enter your keyword(s) here

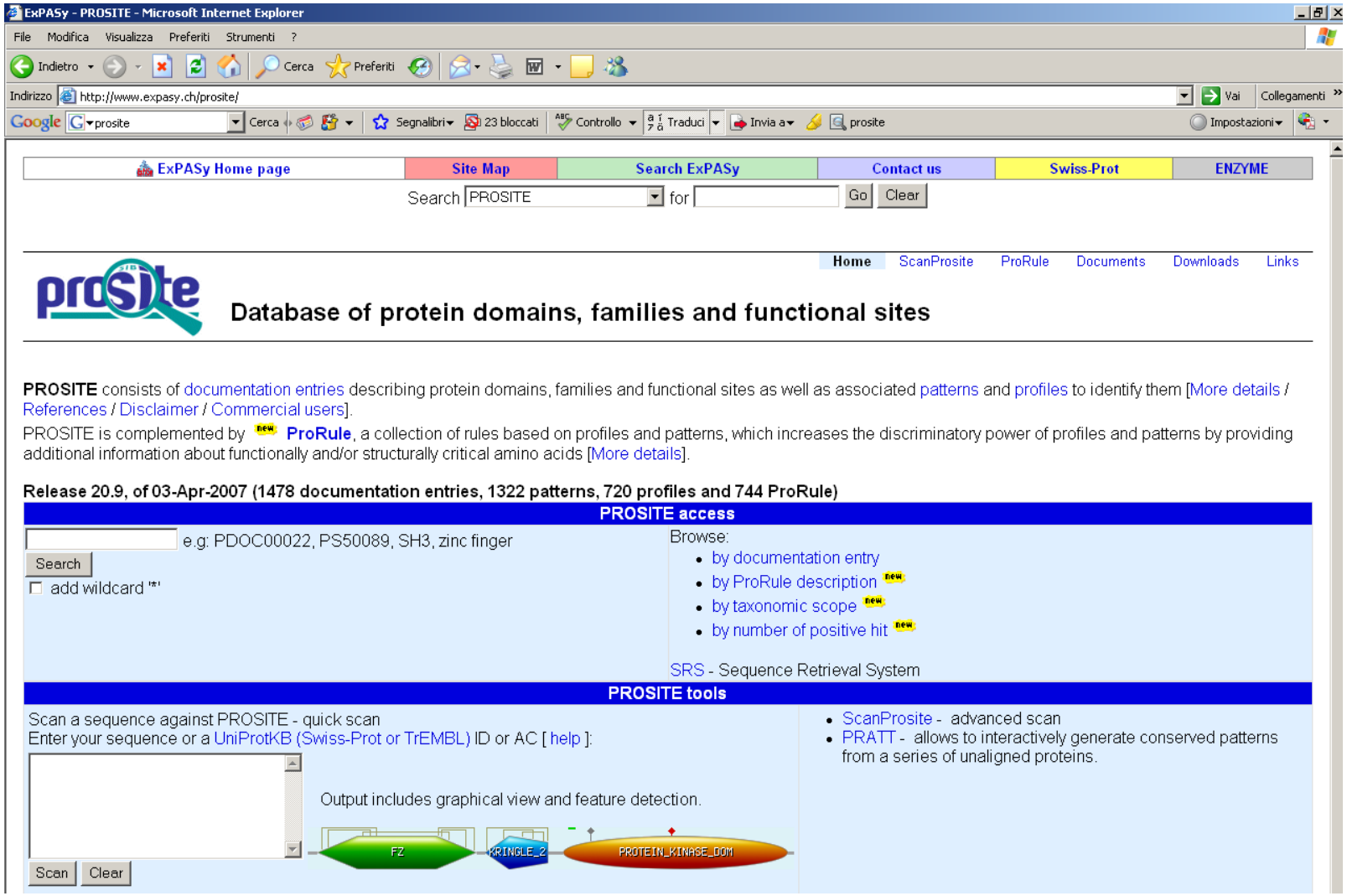
Go Example

Searchable with a novel protein sequence in order to find matches



*Example of a HMM profile fo the **Toxin_7 family**, from the **Pfam database***

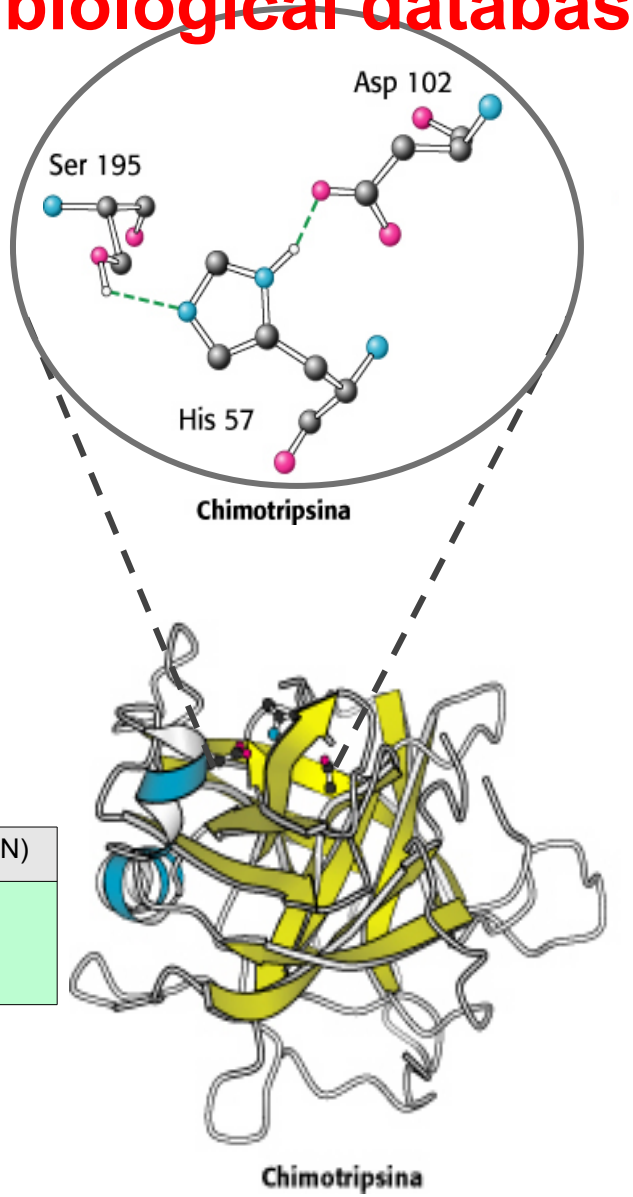
Examples of secondary (derived) biological databases



proSite is a database of protein families, domains and functional sites

Examples of secondary (derived) biological databases

Example of a  pattern

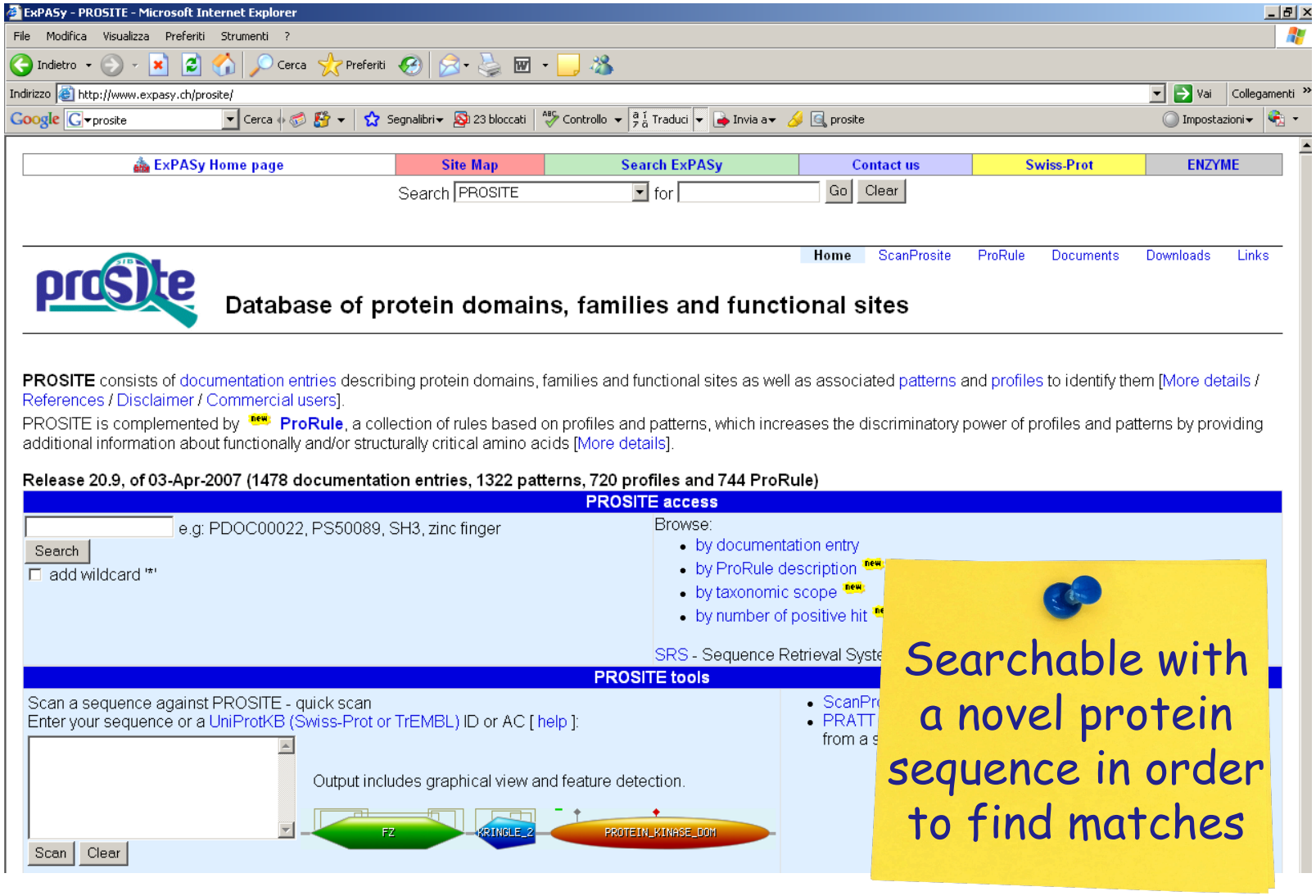


Histidine (H) and Serine (S) belonging to the active site of trypsin-like Ser-protease enzymes appear in sequence patterns of the type reported below

TRYPSIN_HIS, PS00134; **Serine proteases, trypsin family, histidine active site** (PATTERN)	
Consensus pattern:	**[LIVM] - [ST] - A - [STAG] - H - C** *H is the active site residue*

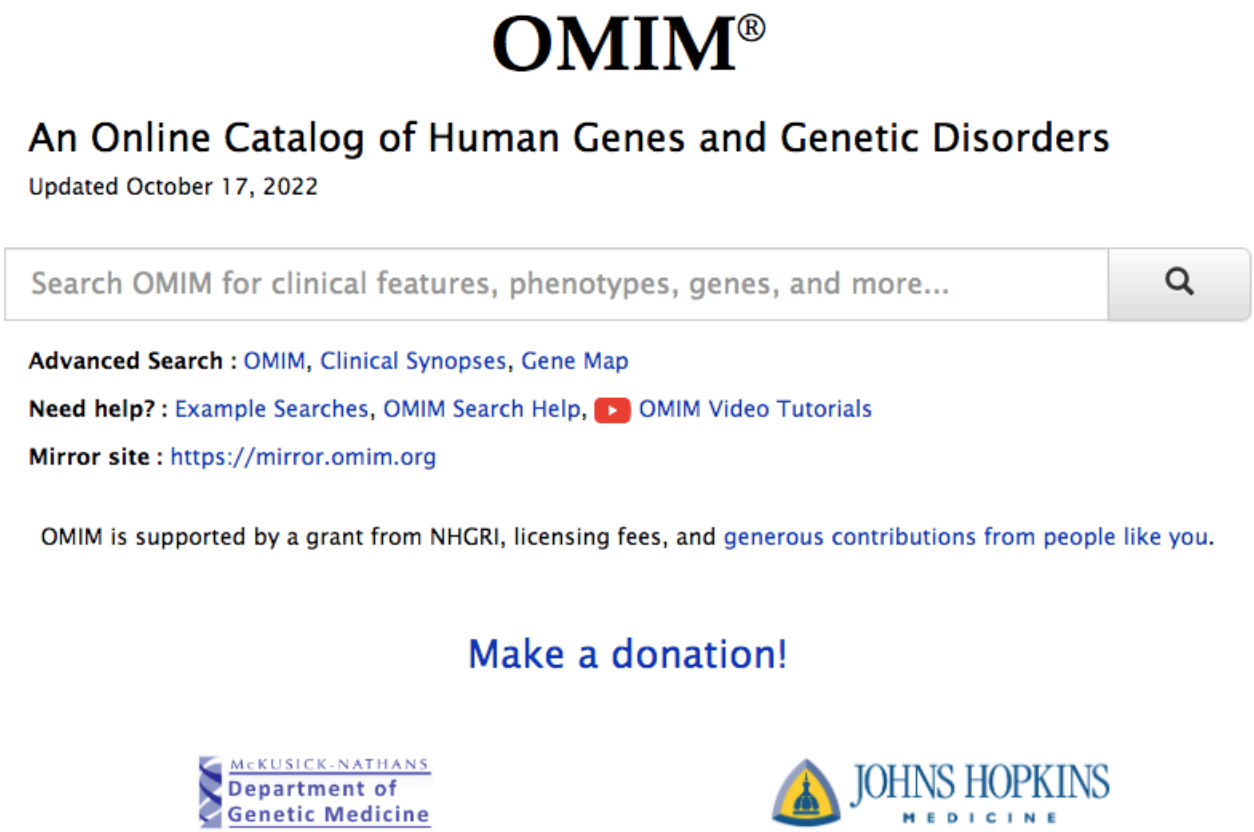
TRYPSIN_SER, PS00135; **Serine proteases, trypsin family, serine active site** (PATTERN)	
Consensus pattern:	**[DNSTAGC] - [GSTAPIMVQH] - x(2) - G - [DE] - S - G - [GS] - [SAPHV] - [LIVMFYWH] - [LIVMFYSTANQH]** *S is the active site residue*

Examples of secondary (derived) biological databases



proSite is a database of protein families, domains and <u>functional sites</u>

Examples of secondary (derived) biological databases



OMIM (Online Mendelian Inheritance in Man) is a catalog of human genes and genetic disorders

Main integrated databases

A number of **centers** were funded to provide access to a large number of major databases in an **integrated environment**

Two main molecular biology database systems exist, which provide integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, homology search software and more

Entrez @ NCBI (National Center for Biotechnology Information) – based in Bethesda, Maryland (USA)

EMBL-EBI (European Bioinfomatics Institute) – based in Hinxton, Cambridge (UK)

Main integrated databases

National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI

hemoglobin

Search

Results found in 30 databases

Literature	
Bookshelf	11,351
MeSH	923
NLM Catalog	1,035
PubMed	239,625
PubMed Central	430,147

Genes	
Gene	8,859
GEO DataSets	4,310
GEO Profiles	292,556
HomoloGene	15
PopSet	239

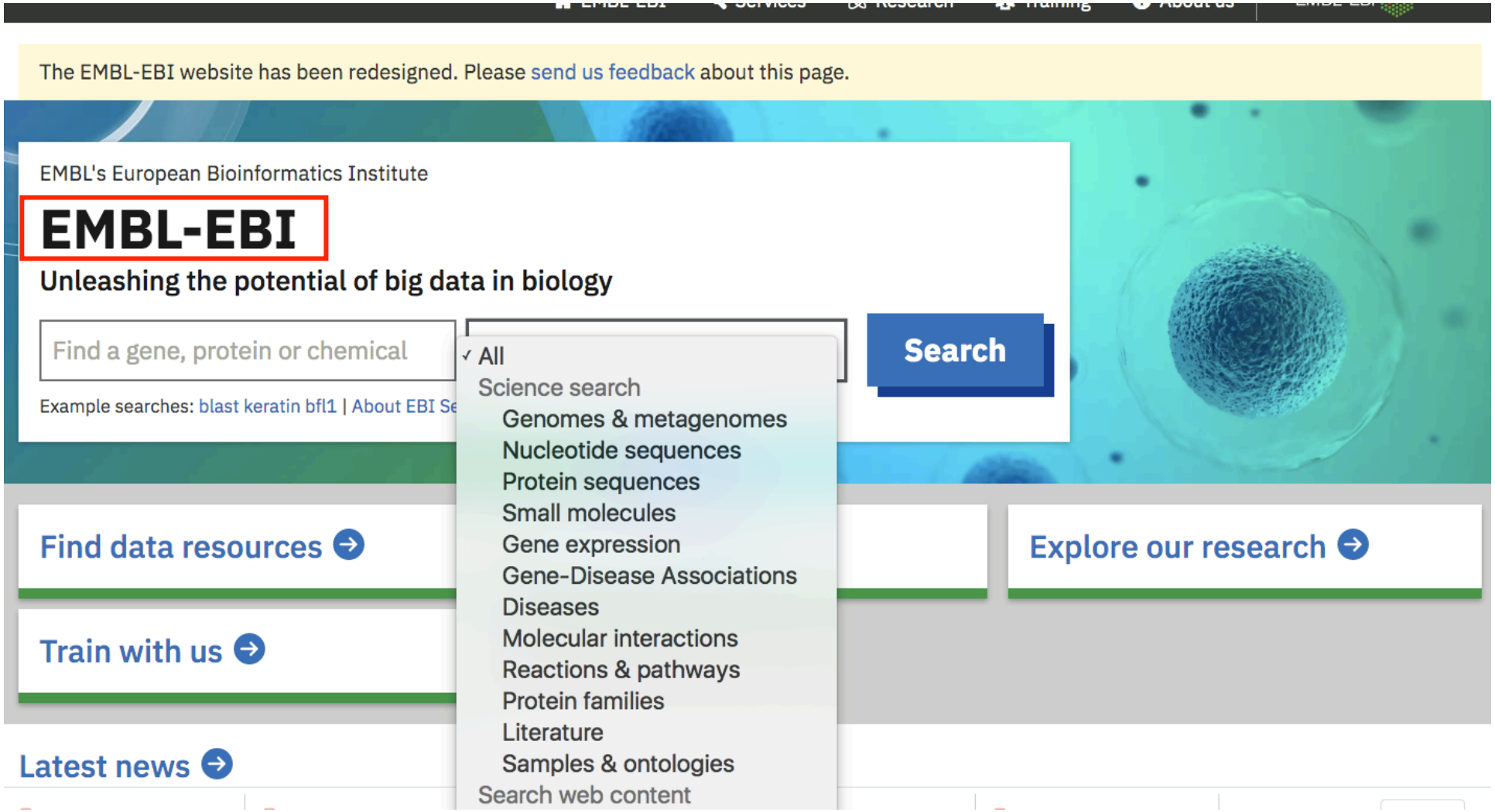
Proteins	
Conserved Domains	35
Identical Protein Groups	51,040
Protein	249,802
Protein Family Models	167
Structure	1,005

Genomes	
Assembly	0
BioCollections	0
BioProject	305
BioSample	2,200
Genome	0
Nucleotide	195,607
SRA	4,553
Taxonomy	0

Clinical	
ClinicalTrials.gov	12,847
ClinVar	2,438
dbGaP	104
dbSNP	0
dbVar	8
GTR	485
MedGen	1,247
OMIM	288

PubChem	
BioAssays	2,303
Compounds	11
Pathways	31
Substances	677

Main integrated databases



Focalizing a DB search….

*Databases usually provide **fields** which can be selected in order to focalize a search, e.g. author ([au]), organism [OS], etc*

Immagine you wish to search a scientific article or a protein sequence by a scientist named **E**(lisabetta) **Coli**

You will have to specify that "**E Coli**" refers to an author (& not a species)!!

e.g. in PubMed: scientific literature database, let's use:

"E coli"

"Coli E[au]"

Example from A. Lesk, "Introduction to Bioinformatics" – Oxford University Press

Edit View Favorites Tools Help

Search Favorites Media

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed

NCBI

PubMed www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for E Coli Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Send to

All: 200777 Review: 7313

Items 1 - 20 of 200777 Page 1 of 10039 Next

About Entrez
NCBI Toolbar

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway

1: Kim SM, Kuzuyama T, Chang YJ, Kim SU. Links

Cloning and characterization of 2-C-methyl-D: -erythritol 2,4-cyclodiphosphate synthase (MECS) gene from Ginkgo biloba.
Plant Cell Rep. 2006 Mar 10; [Epub ahead of print]
PMID: 16528563 [PubMed - as supplied by publisher]

2: Zelic B, Bolf N, Vasic-Racki D. Links

Modeling of the pyruvate production with Escherichia coli: comparison of mechanistic and neural networks-based models.
Bioprocess Biosyst Eng. 2006 Mar 10; [Epub ahead of print]
PMID: 16528533 [PubMed - as supplied by publisher]

3: Lahiri C, Mandal S, Ghosh W, Dam B, Roy P. Links

A Novel Gene Cluster soxSRT Is Essential for the Chemolithotrophic Oxidation of Thiosulfate and Tetrathionate by Pseudaminobacter salicylatoxidans KCT001.
Curr Microbiol. 2006 Mar 9; [Epub ahead of print]
PMID: 16528465 [PubMed - as supplied by publisher]

4: Cookson AL, Taylor SC, Attwood GT. Links

The prevalence of Shiga toxin-producing Escherichia coli in cattle and sheep in the lower North Island, New Zealand

Internet

Edit View Favorites Tools Help

Search Favorites Media

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed

NCBI

PubMed www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for Coli E[au] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Send to

All: 9 Review: 1

Items 1 - 9 of 9 One page

1: Ramacciotti CE, Coli E, Paoli R, Gabriellini G, Schulte F, Castrogiovanni S, Dell'Osso L, Garfinkel PE. Related Articles, Links

The relationship between binge eating disorder and non-purging bulimia nervosa.
Eat Weight Disord. 2005 Mar;10(1):8-12.
PMID: 15943166 [PubMed - in process]

2: Ramacciotti CE, Guidi L, Bondi E, Coli E, Dell'Osso L, Pistoia S, Pucci E. Related Articles, Links

Differential dynamic responses of luteinizing hormone to gonadotropin releasing hormone in patients affected by bulimia nervosa-purging versus non-purging type.
Eat Weight Disord. 1997 Sep;2(3):150-5.
PMID: 14655839 [PubMed - indexed for MEDLINE]

3: Lenzi A, Maltinti E, Poggi E, Fabrizio L, Coli E. Related Articles, Links

Effects of rivastigmine on cognitive function and quality of life in patients with schizophrenia.
Clin Neuropharmacol. 2003 Nov-Dec;26(6):317-21.
PMID: 14646612 [PubMed - indexed for MEDLINE]

4: Ramacciotti CE, Coli E, Biadi O, Dell'Osso L. Related Articles, Links

Silent pericardial effusion in a sample of anorexic patients.
Eat Weight Disord. 2003 Mar;8(1):68-71.
PMID: 12762627 [PubMed - indexed for MEDLINE]

5: Ramacciotti CE, Coli E, Paoli R, Marazziti D, Dell'Osso L. Related Articles, Links

About Entrez
NCBI Toolbar

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway

Internet

Quality of data in DBs

Data and relative annotations in DBs may contain **errors**, so the **accuracy** can be an issue

Some of these errors can be checked automatically, by a computer-based approach (*e.g. that DNA sequences only consist of A/G/T/C*)

Some other errors can only be checked by manual curation

Quality of data in DBs

In addition, databases can be **redundant**, which means that several entries can have identical data

Ideally, all these entries with same data should be included in a single entry, which refers to all the independent experiments and summarizes all the separate DB entries

In this way, a **nonredundant database** is obtained, which allows to discover all the information by reading the single entry

Quality of data in DBs

Databases need to be regularly updated:

- to include new entries and
- to update and correct existing ones

It is important to recognize if the current entry differs from a copy made earlier, this can be done by using version numbers for entries or reporting the most recent date on which changes were made. A **unique identifier** for the entry is needed!

Entries in DBs usually have a maturation

Not annotated → Preliminary → Not reviewed → Standard

For instance, in early stages of a genome annotation, genes are identified by only applying computational methods; they are thus annotated as "hypothetical" genes (translated proteins as hypothetical proteins)

Then, when experimental evidence becomes available confirming or not a gene, the relative entry will be updated

Updated list of available and relevant biological DBs

OXFORD ACADEMIC | Journals

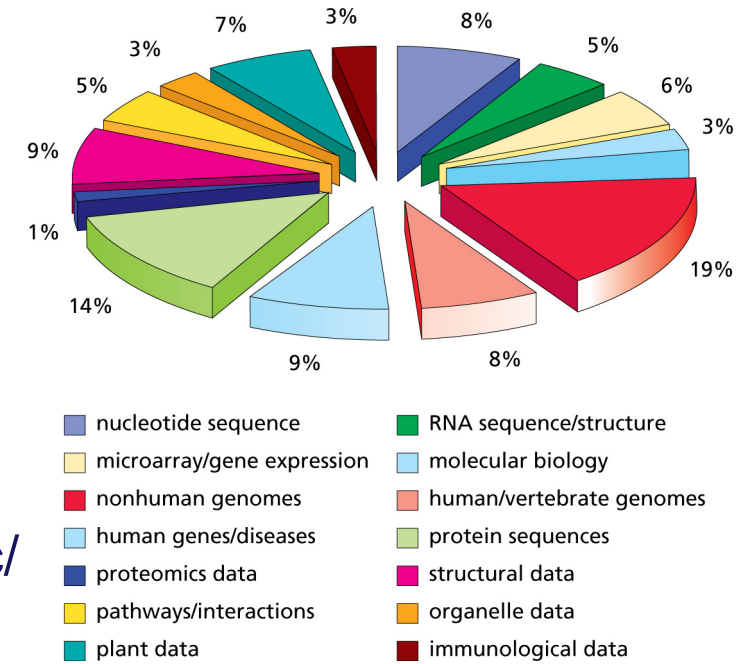
Updated every January on Nucleic Acids Research (NAR) – Oxford Journals

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

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

NAR Database Summary Paper Category List

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



statistics not updated

https://www.oxfordjournals.org/nar/database/c/

Lesson 7.
Content

1. Biological databases. Constitute the backbone of bioinformatics research. Can be primary or secondary. Largely public and extremely useful although they may contain and propagate errors