Biological Data Sources and File Formats

iCB2 – Introduction to Computational Biology and Bioinformatics
November 12, 2015

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Institute for Mathematical Modeling
of Biological Systems
Department of Biology

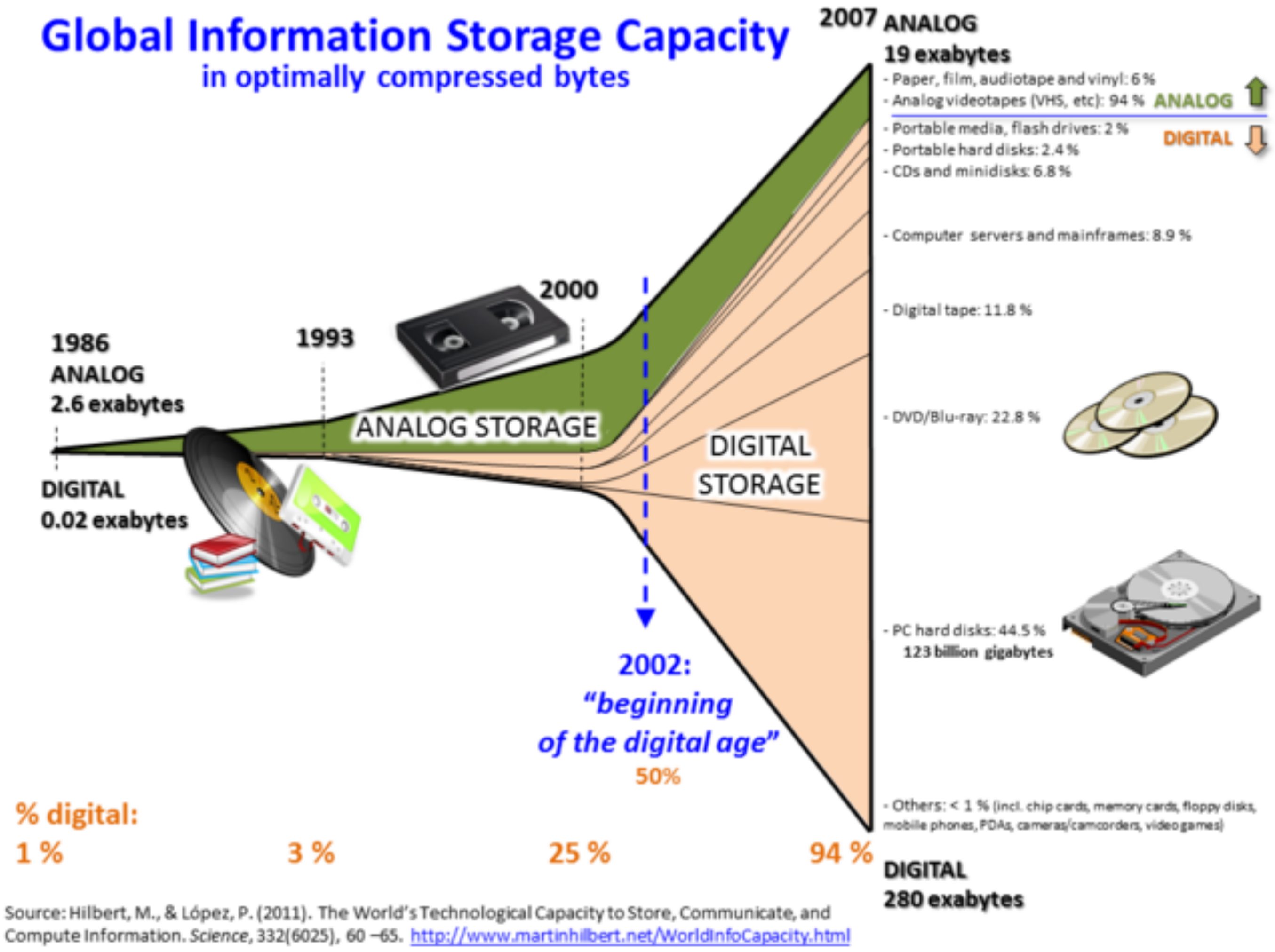
HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Big Data

Big Data refers to data sets so large or complex that they are difficult to process using traditional data processing applications.

Main challenges include:

- analysis
- capture
- curation
- search
- sharing
- storage
- transfer
- visualization
- information privacy.

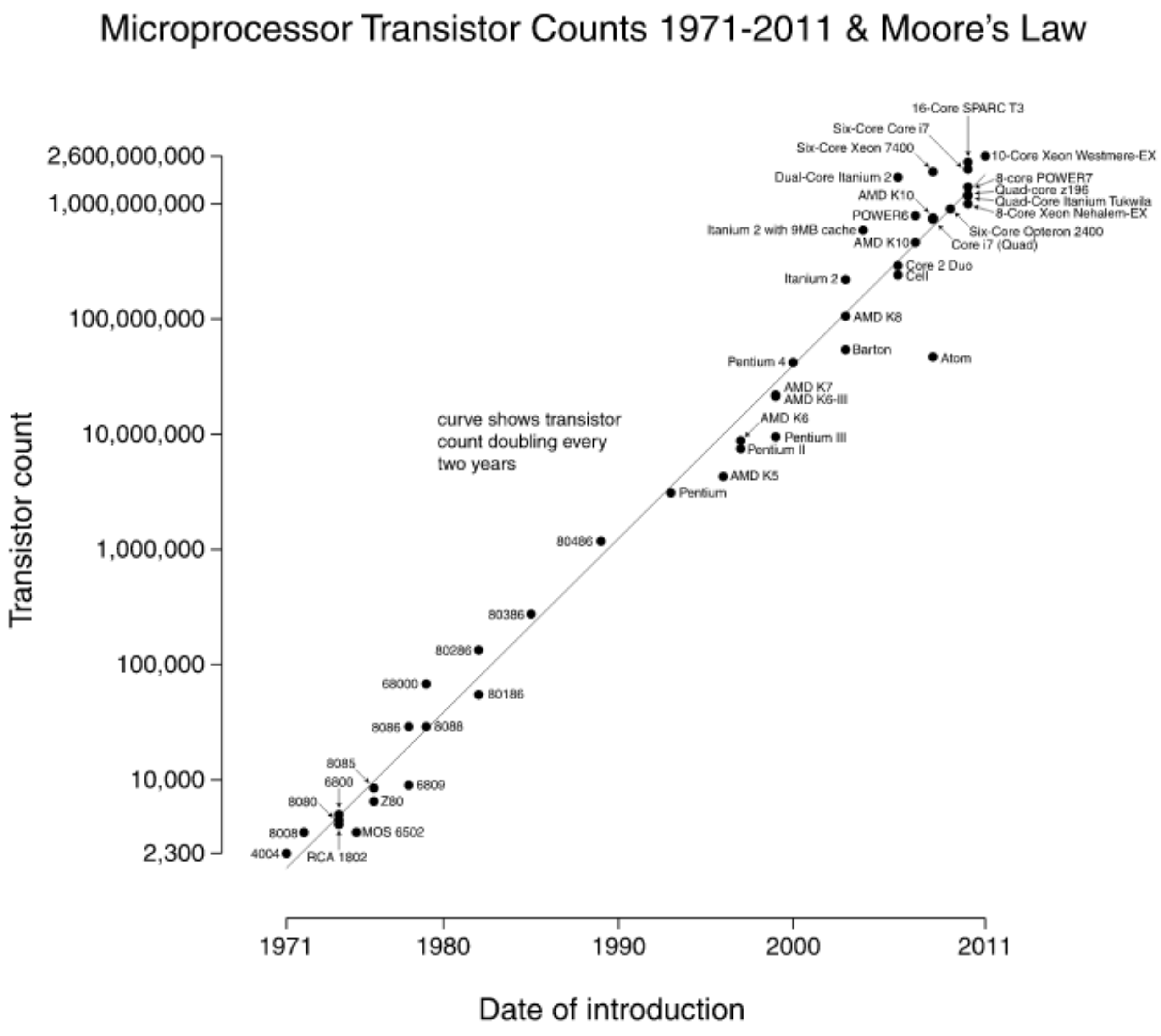


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

from wikipedia

Moore’s Law

It is based on the observation that, over the history of computing hardware, the number of transistors in a dense integrated circuit doubles approximately every two years.



from wikipedia

Big Data in biology

The complete human genome in the 2004 was released in 2004

International HGS Consortium Nature 2004. PMID: 15496913

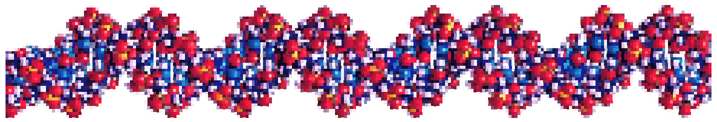
International consortiums such as HapMap, 1000Genomes and ENCODE are collecting large amount of data about the human genome.

The NCBI collects the complete genomic sequences of many organisms

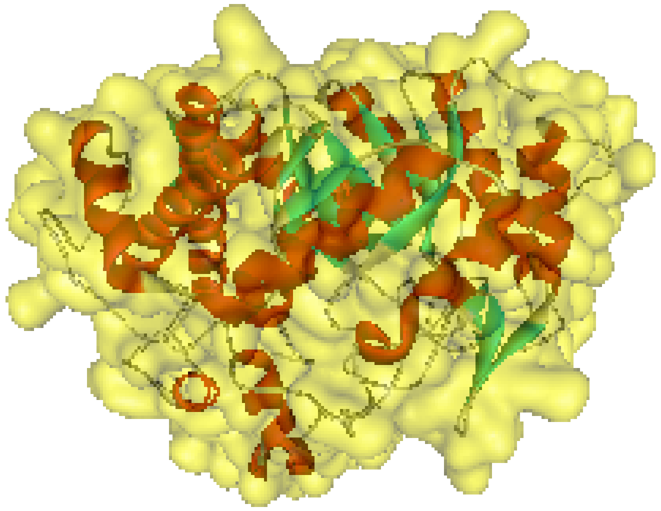
- Archea: 212/605 species
- Bacteria: 4,903/53,392 species
- Eukariots: 346/2,423 species

Nov 2015

Molecular biology data



```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```



GenBank:	188,372,017
UniRef90:	36,805,263
Swiss-Prot:	549,832
Protein Data Bank:	113,672
Protein:	105,572
Nucleic Acids:	2,859

Nov 2015

Definition

Computational Biology and Bioinformatics: the same focus but different priorities

Bioinformatics is an interdisciplinary field that **develops methods and software** tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data.

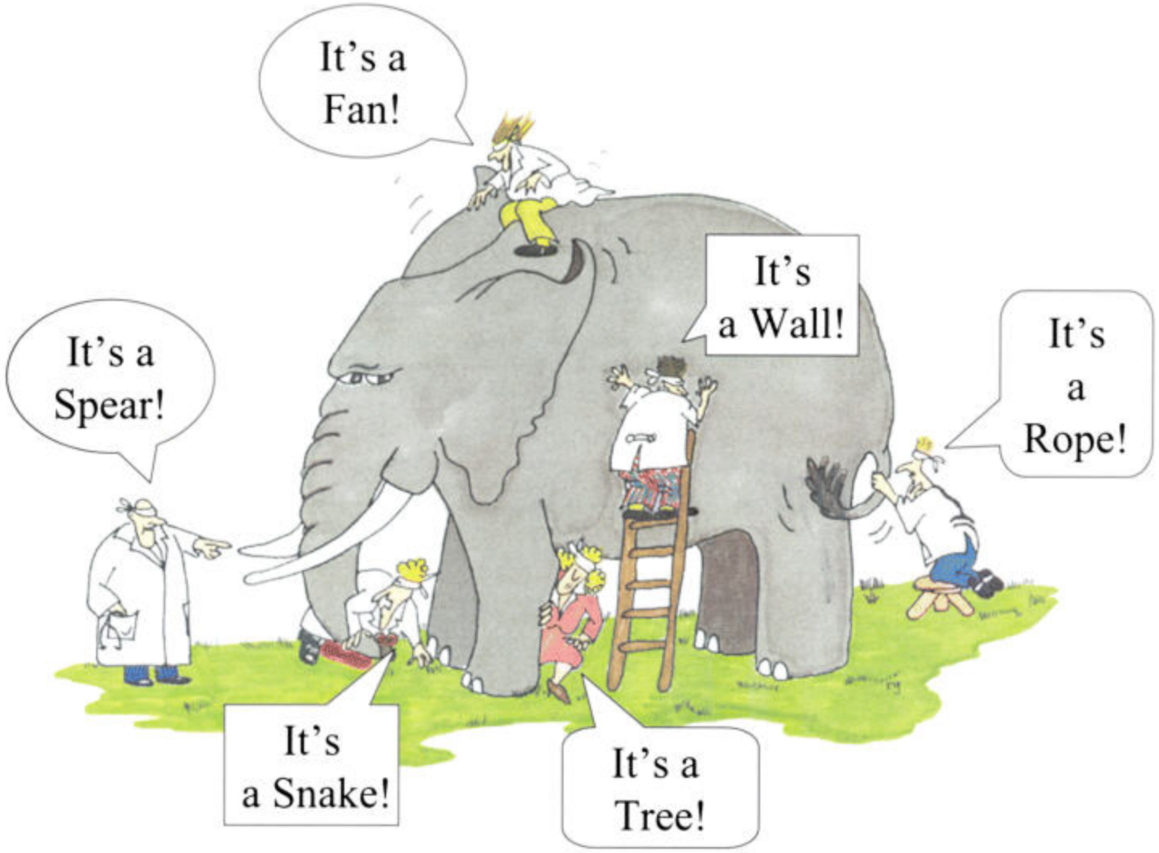
Computational biology involves the development and **application of data-analytical and theoretical methods**, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

from wikipedia

The elephant or the cave?

Computer sciences: building the big eye of the future

Scientists and the elephant



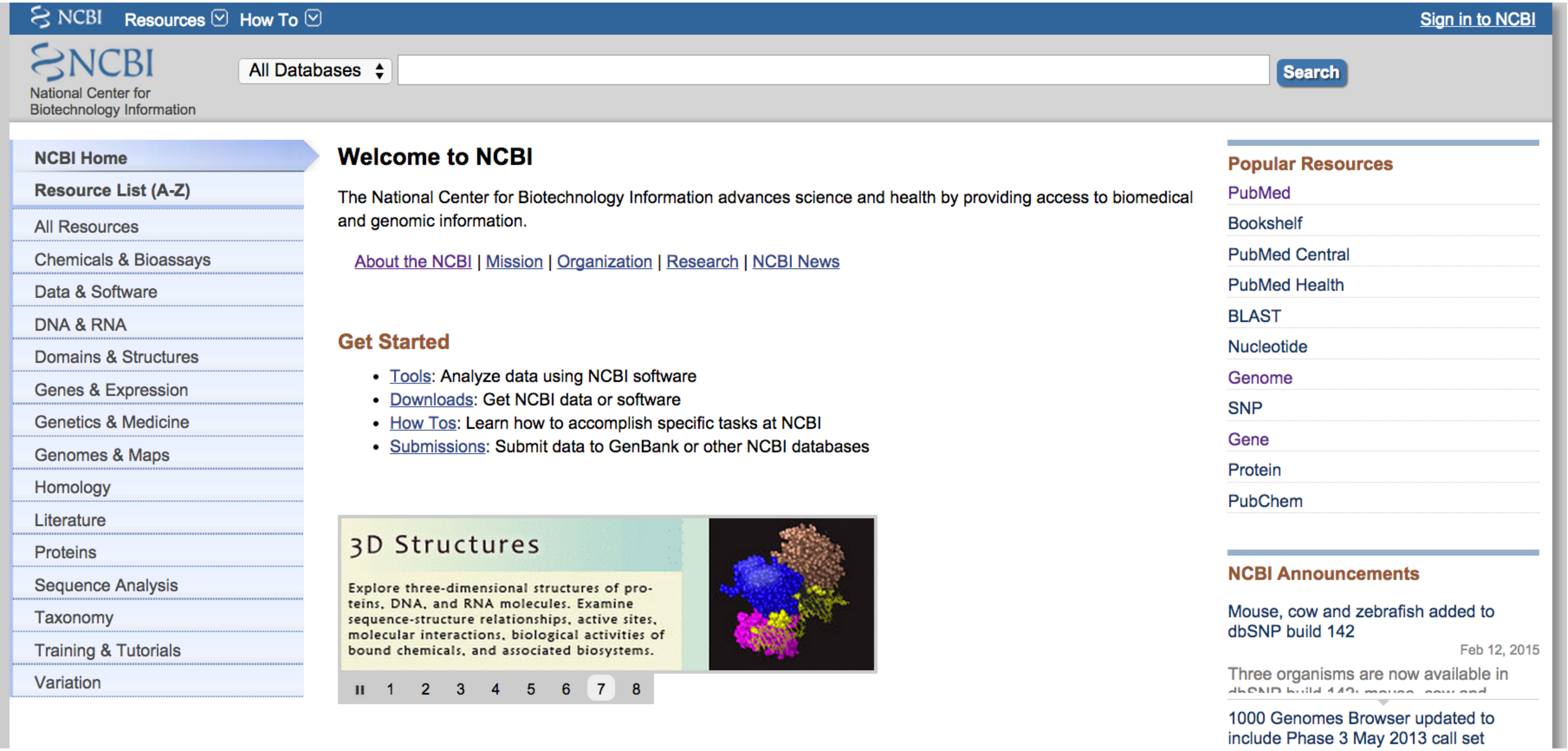
www.goldismoney2.com

Plato's Allegory of the Cave (*The Republic*)

www.dailyimpact.net

The NCBI

Many resources and primary databases with molecular biology data.
Some examples are GenBank, RefSeq, GEO, dbSNP, dbGAP .....



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

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

II 1 2 3 4 5 6 7 8

Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements

Mouse, cow and zebrafish added to dbSNP build 142

Feb 12, 2015

Three organisms are now available in dbSNP build 142: mouse, cow and

1000 Genomes Browser updated to include Phase 3 May 2013 call set

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

Main data types

In molecular biology several type of data are available. Among the most common there are:

- Sequences: string representing the nucleotide and amino acid composition of DNA, RNA and protein.

- Annotations: collection of words with controlled vocabulary that describes property, function, and process in which a biomolecule is involved.

- Structure: 2D or 3D representation of a molecule describing how it it is organized in the space.

The Sequence

Most common format is FASTA, which is a text file containing an header starting with “>” and a single or multiple lines of strings representing the nucleotides of the amino acids in one letter codes.

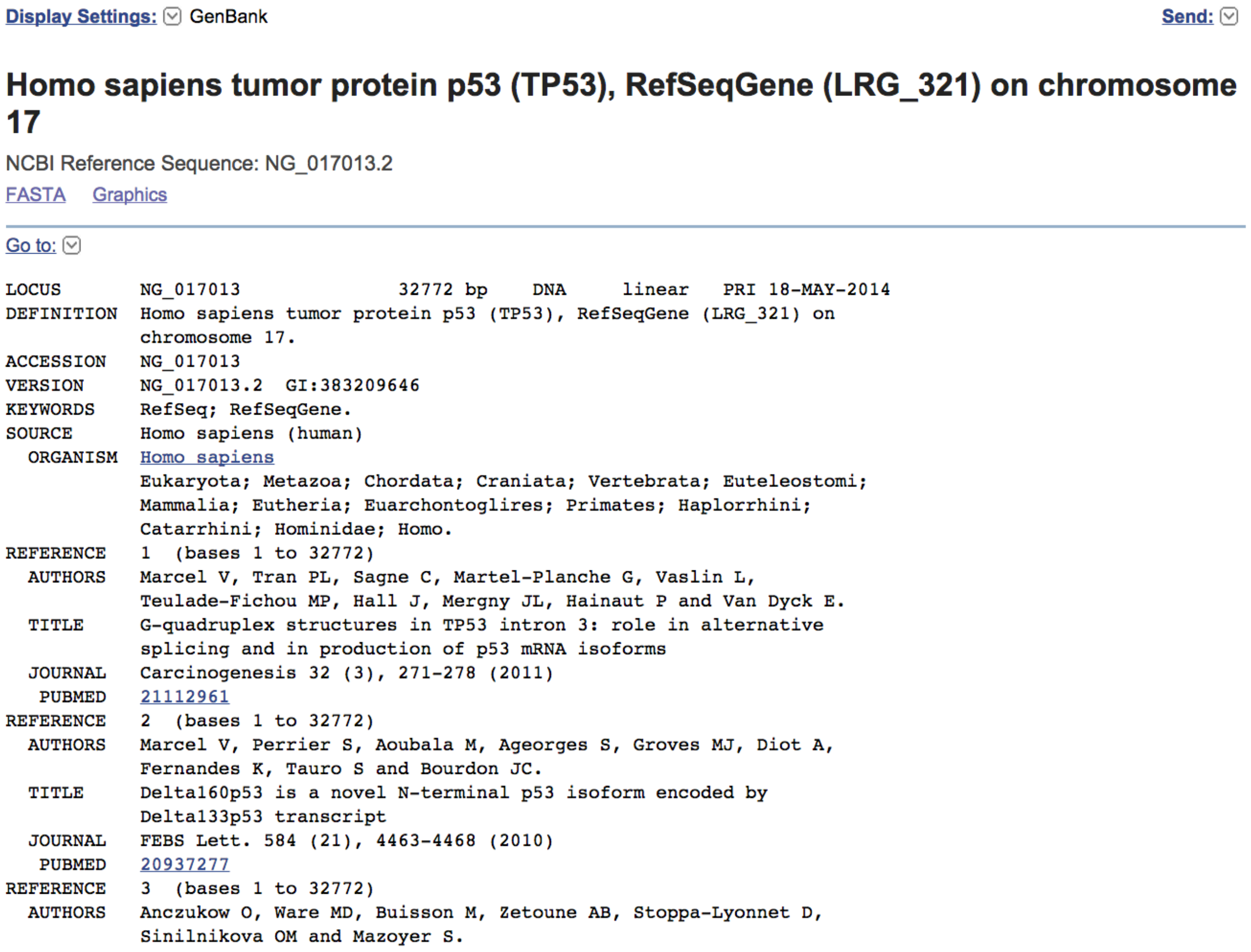
```
>ref|NG_017013.2| Homo sapiens tumor protein p53 (TP53)
CTCCTTGGTTCAAGTAATTCTCCTGCCTCAGACTCCAGAGTAGCTGGGATTACAGGCGCCCGCCACCACG
CCCAGCTAATTTTTTGTATTTTTAATAGAGATGGGGTTTCATCATGTTGGCCAGGCTGGTCTCGAACTCC
TGACCTCAGGTGATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAGGAGTCAGCCACCGCACCCA
......
```

Another old time sequence format is the PIR (Protein Information Resource)

```
>P1;CRAB_ANAPL
ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDITIHNPLIRRPLFSWLAPSRIFDQIFGEHLQESELLPASPSLSPFLMRSPIFRMPSWLETGLSEMRLEK
DKFSVNLDVKHFSPEELKVKVLGDMVEIHGKHEERQDEHGFIAREFNRKYRIPADVDPLTITSSLSLDGVL
TVSAPRKQSDVPERSIPITREEKPAIAGAQRK*
```

GenBank

Is the most comprehensive database of DNA sequences from several organisms. Sequence are associated to a Gene Identifier (GI).



Display Settings: GenBank   Send:

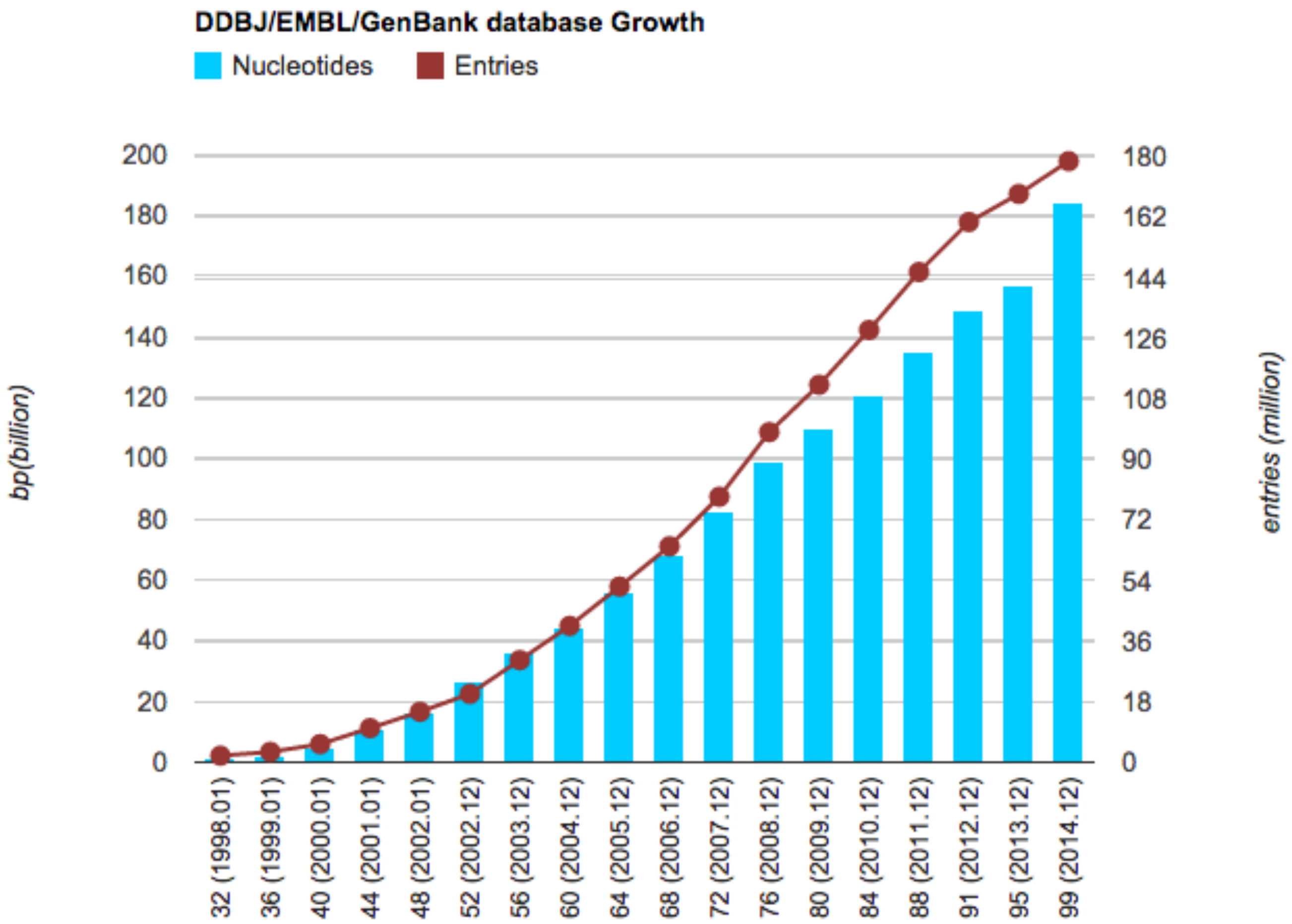
Homo sapiens tumor protein p53 (TP53), RefSeqGene (LRG_321) on chromosome 17

NCBI Reference Sequence: NG_017013.2

FASTA Graphics

Go to:

```
LOCUS       NG_017013              32772 bp    DNA     linear   PRI 18-MAY-2014
DEFINITION  Homo sapiens tumor protein p53 (TP53), RefSeqGene (LRG_321) on
            chromosome 17.
ACCESSION   NG_017013
VERSION     NG_017013.2  GI:383209646
KEYWORDS    RefSeq; RefSeqGene.
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 32772)
  AUTHORS   Marcel V, Tran PL, Sagne C, Martel-Planche G, Vaslin L,
            Teulade-Fichou MP, Hall J, Mergny JL, Hainaut P and Van Dyck E.
  TITLE     G-quadruplex structures in TP53 intron 3: role in alternative
            splicing and in production of p53 mRNA isoforms
  JOURNAL   Carcinogenesis 32 (3), 271-278 (2011)
   PUBMED   21112961
REFERENCE   2  (bases 1 to 32772)
  AUTHORS   Marcel V, Perrier S, Aoubala M, Ageorges S, Groves MJ, Diot A,
            Fernandes K, Tauro S and Bourdon JC.
  TITLE     Delta160p53 is a novel N-terminal p53 isoform encoded by
            Delta133p53 transcript
  JOURNAL   FEBS Lett. 584 (21), 4463-4468 (2010)
   PUBMED   20937277
REFERENCE   3  (bases 1 to 32772)
  AUTHORS   Anczukow O, Ware MD, Buisson M, Zetoune AB, Stoppa-Lyonnet D,
            Sinilnikova OM and Mazoyer S.
```



GenBank and RefSeq

In GenBank you can have all available versions for each genomic sequence.

Sequences are also indicated with the following codes: NC (chromosomes), NM (mRNAs), NP (proteins), or NT (constructed genomic contigs) and NG (genomic regions or gene clusters)

RefSeq is an annotated and curated dataset that contains a single record for each nucleotide sequences (DNA, RNA) and their protein products.

It is possible to download sequences in using eutils tools

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=**code**&rettype=fasta&retmode=text

TP53: 383209646 or NG_017013

The Annotation

Is the process of assigning to any sequence the features that defines the function and of a nucleotide and protein sequence.

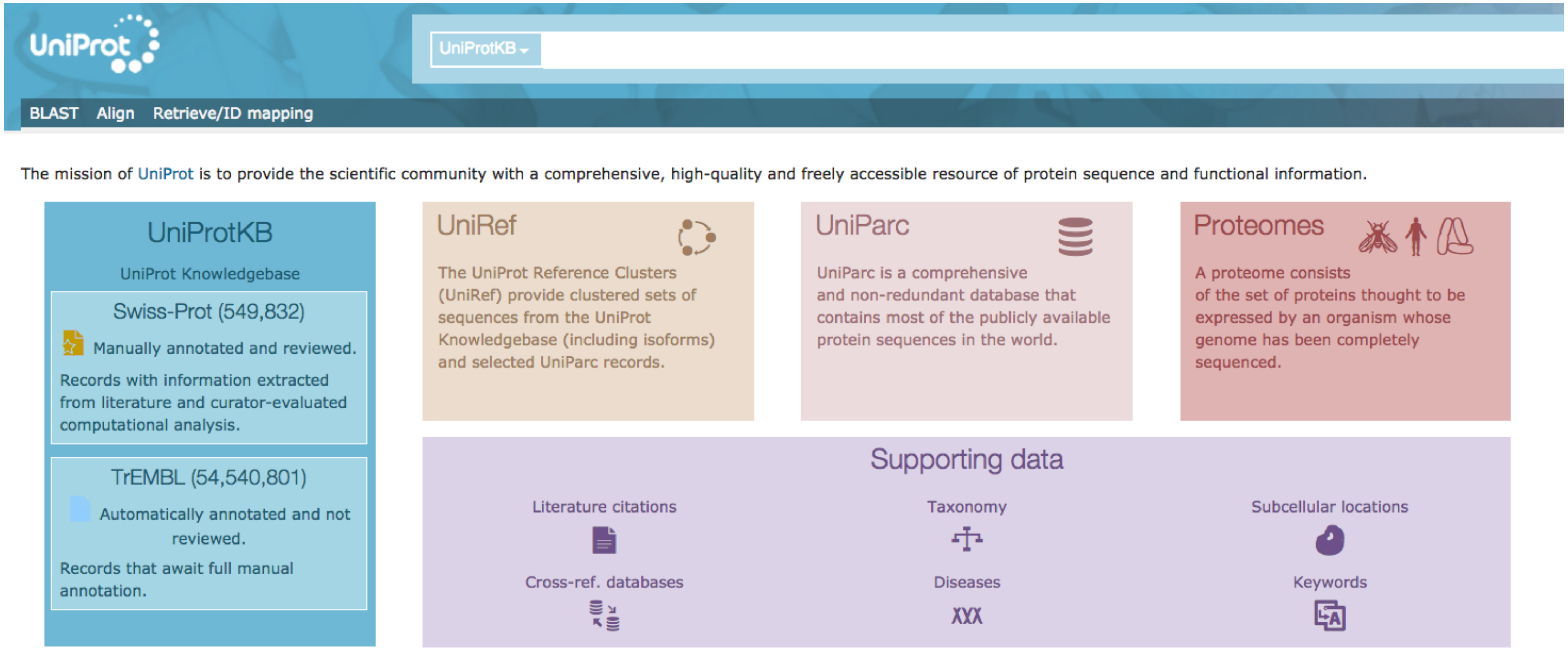
The annotation can be wither either automatic, using computational tools or manual, using results of experimental.

The automatic annotation is mainly based on homology search because

higher sequence similarity => higher the probability similarity in function

The UniProt

The European repository of molecular biology data. UniProtKB is composed by SwissProt and TrEMBL



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

UniProtKB

UniProt Knowledgebase

Swiss-Prot (549,832)

Manually annotated and reviewed.

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

TrEMBL (54,540,801)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

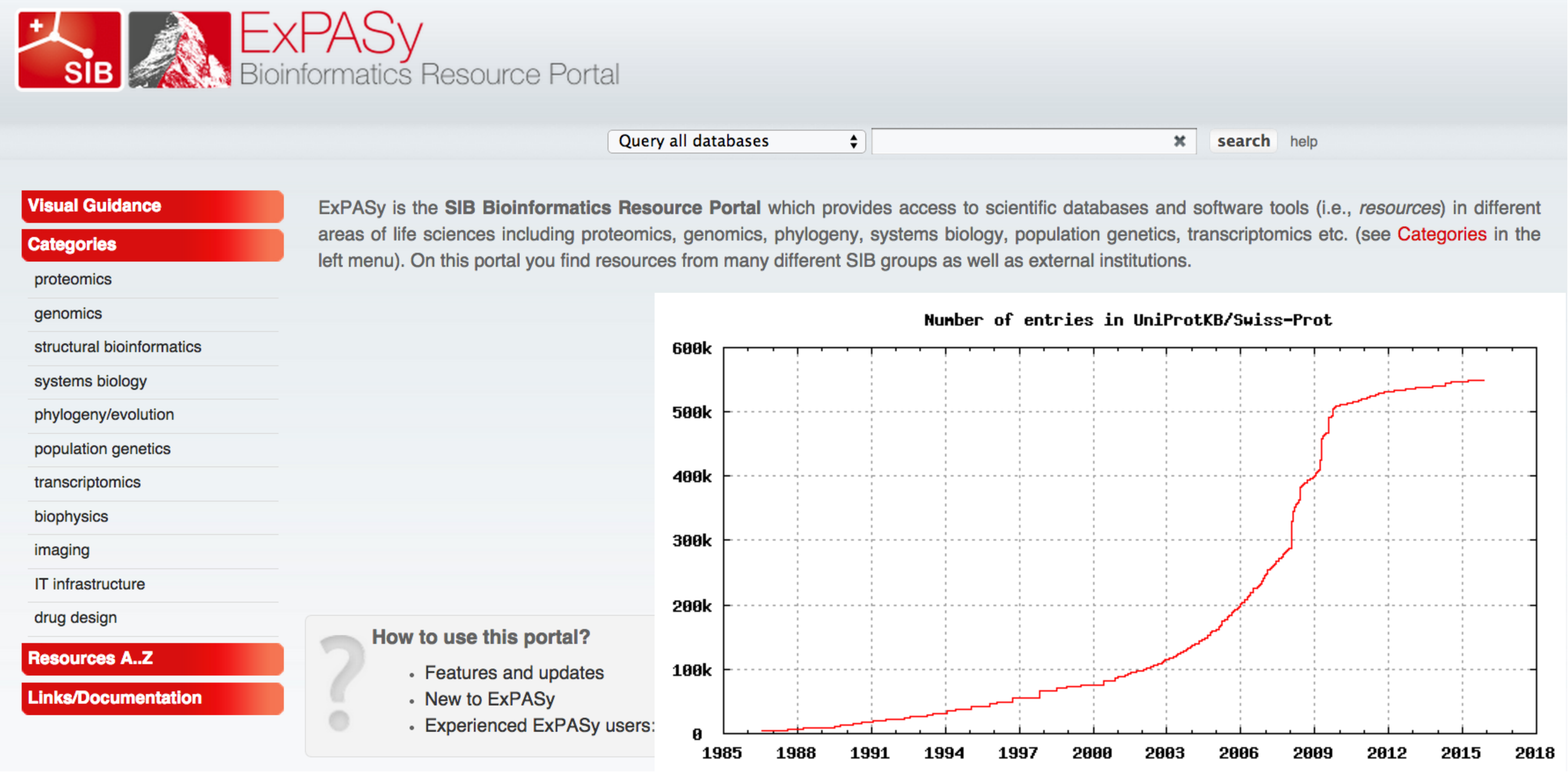
A proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

Supporting data

http://www.uniprot.org/

The SwissProt

SwissProt contains all the proteins that have been manually annotated using information extracted from literature.



ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

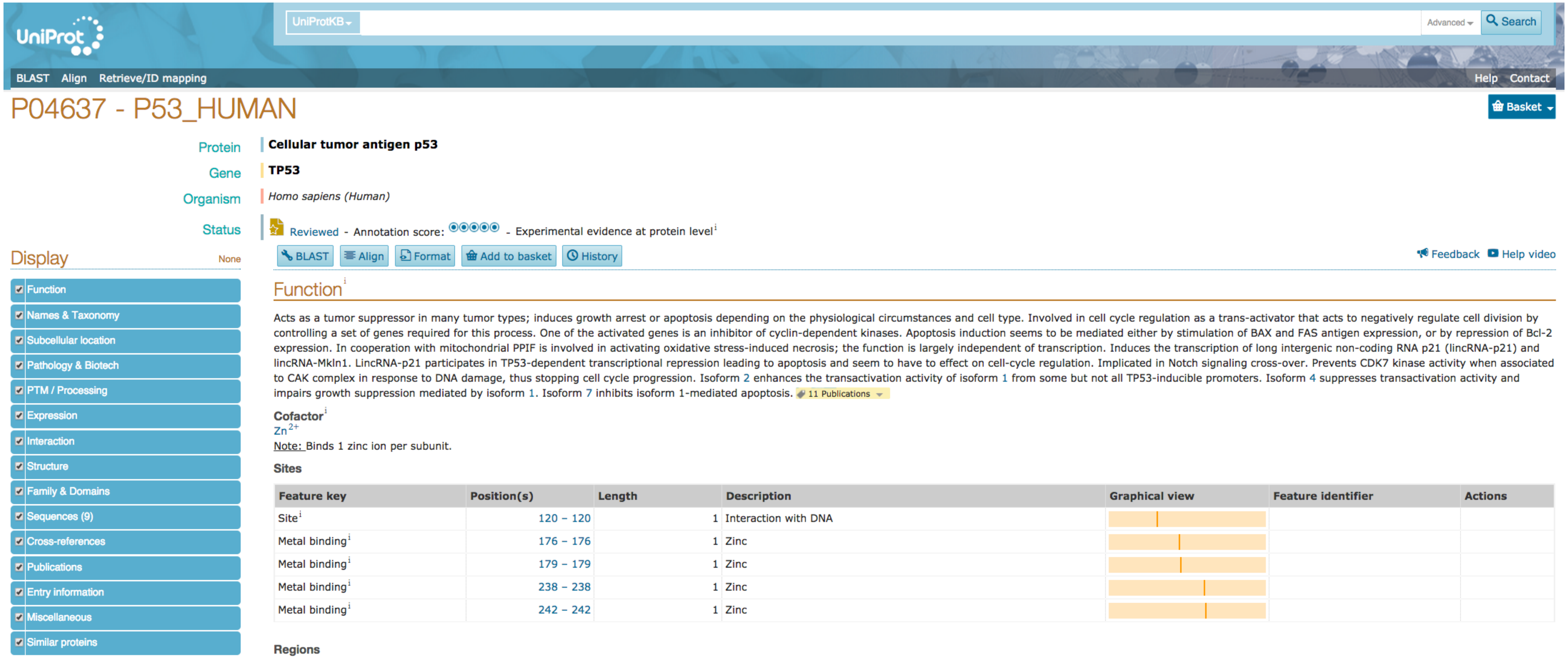
How to use this portal?

- Features and updates
- New to ExPASy
- Experienced ExPASy users:

http://www.expasy.org/

The function

Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type.
11 Publication



P04637 - P53_HUMAN

Protein | **Cellular tumor antigen p53**

Gene | **TP53**

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PPIF is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lincRNA-p21) and lincRNA-Mkln1. LincRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have to effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1. Isoform 7 inhibits isoform 1-mediated apoptosis. 11 Publications

Cofactor
Zn^{2+}
Note: Binds 1 zinc ion per subunit.

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Site	120 – 120	1	Interaction with DNA			
Metal binding	176 – 176	1	Zinc			
Metal binding	179 – 179	1	Zinc			
Metal binding	238 – 238	1	Zinc			
Metal binding	242 – 242	1	Zinc			

Regions

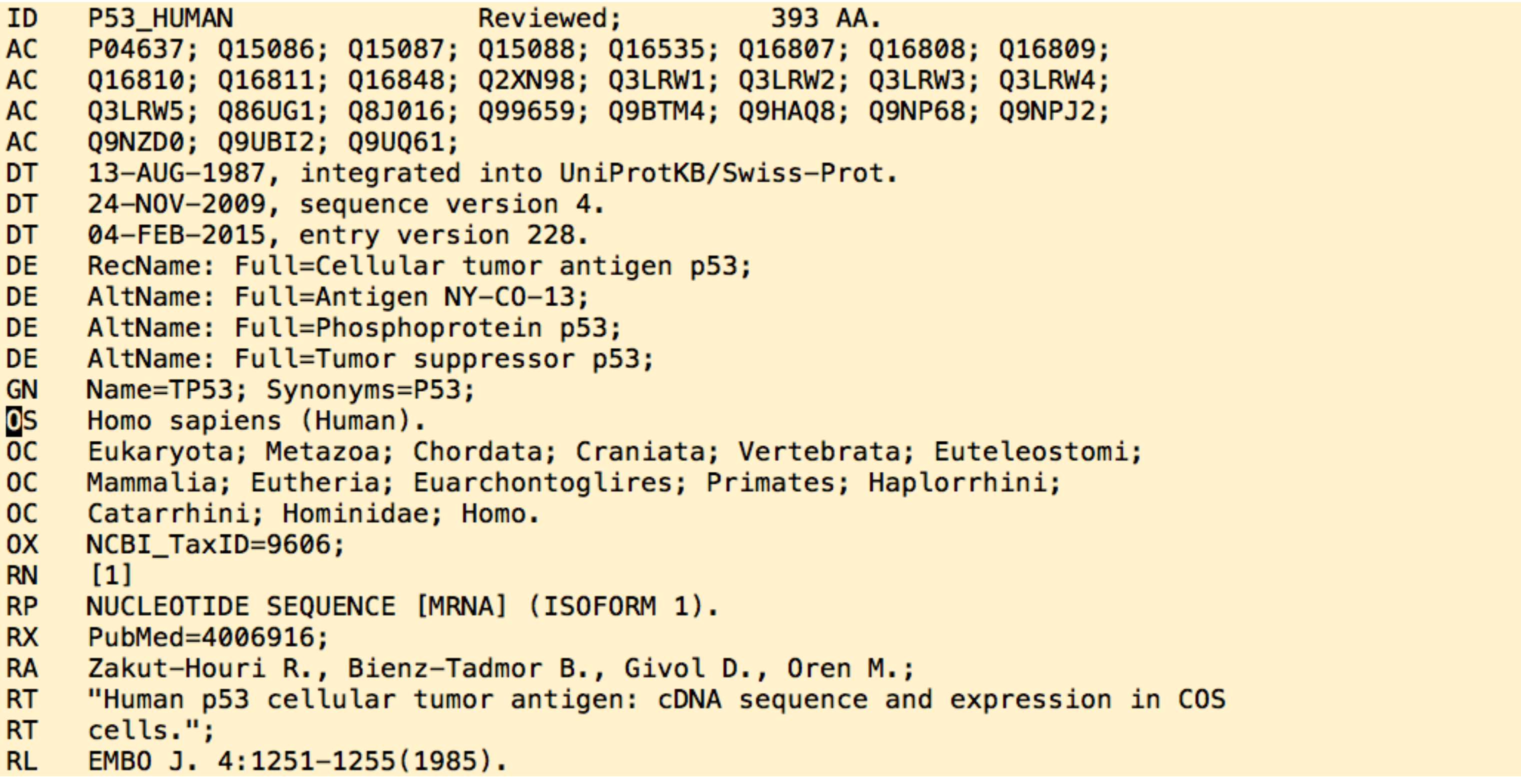
Getting the information

The SwissProt fasta file contains all the sequences in the database and the dat file contains all the information including annotation.

The fasta and dat files can be downloaded using the following links

http://www.uniprot.org/uniprot/P53_HUMAN.fasta
http://www.uniprot.org/uniprot/P53_HUMAN.txt

More complex queries:
http://www.uniprot.org/help/programmatic_access



```
ID   P53_HUMAN               Reviewed;         393 AA.
AC   P04637; Q15086; Q15087; Q15088; Q16535; Q16807; Q16808; Q16809;
AC   Q16810; Q16811; Q16848; Q2XN98; Q3LRW1; Q3LRW2; Q3LRW3; Q3LRW4;
AC   Q3LRW5; Q86UG1; Q8J016; Q99659; Q9BTM4; Q9HAQ8; Q9NP68; Q9NPJ2;
AC   Q9NZD0; Q9UBI2; Q9UQ61;
DT   13-AUG-1987, integrated into UniProtKB/Swiss-Prot.
DT   24-NOV-2009, sequence version 4.
DT   04-FEB-2015, entry version 228.
DE   RecName: Full=Cellular tumor antigen p53;
DE   AltName: Full=Antigen NY-CO-13;
DE   AltName: Full=Phosphoprotein p53;
DE   AltName: Full=Tumor suppressor p53;
GN   Name=TP53; Synonyms=P53;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX   PubMed=4006916;
RA   Zakut-Houri R., Bienz-Tadmor B., Givol D., Oren M.;
RT   "Human p53 cellular tumor antigen: cDNA sequence and expression in COS
RT   cells.";
RL   EMBO J. 4:1251-1255(1985).
```

Problem 1.a

Bert Voglestein in a Science paper published in 2013 (PMID: 23539594) reported a list of Tumor Suppressor genes and Oncogenes.

Take the list of Tumor suppressor gene ids and map them to SwissProt ids

1. Download a list of genes from http://biofold.org/courses/docs/vogelstein_tsg.txt

2. Write a bash script to transform the gene id to SwissProt id using the UniProt REST API:

   http://www.uniprot.org/uniprot/?query=organism:9606+AND+gene:GeneID&format=tab&columns=id

Problem 1.b

Write an efficient python script that extracts from the SwissProt fasta file the subset of sequences with Swiss Ids provided in a file list.

1. Download the whole SwissProt database form ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot_sprot.fasta.gz

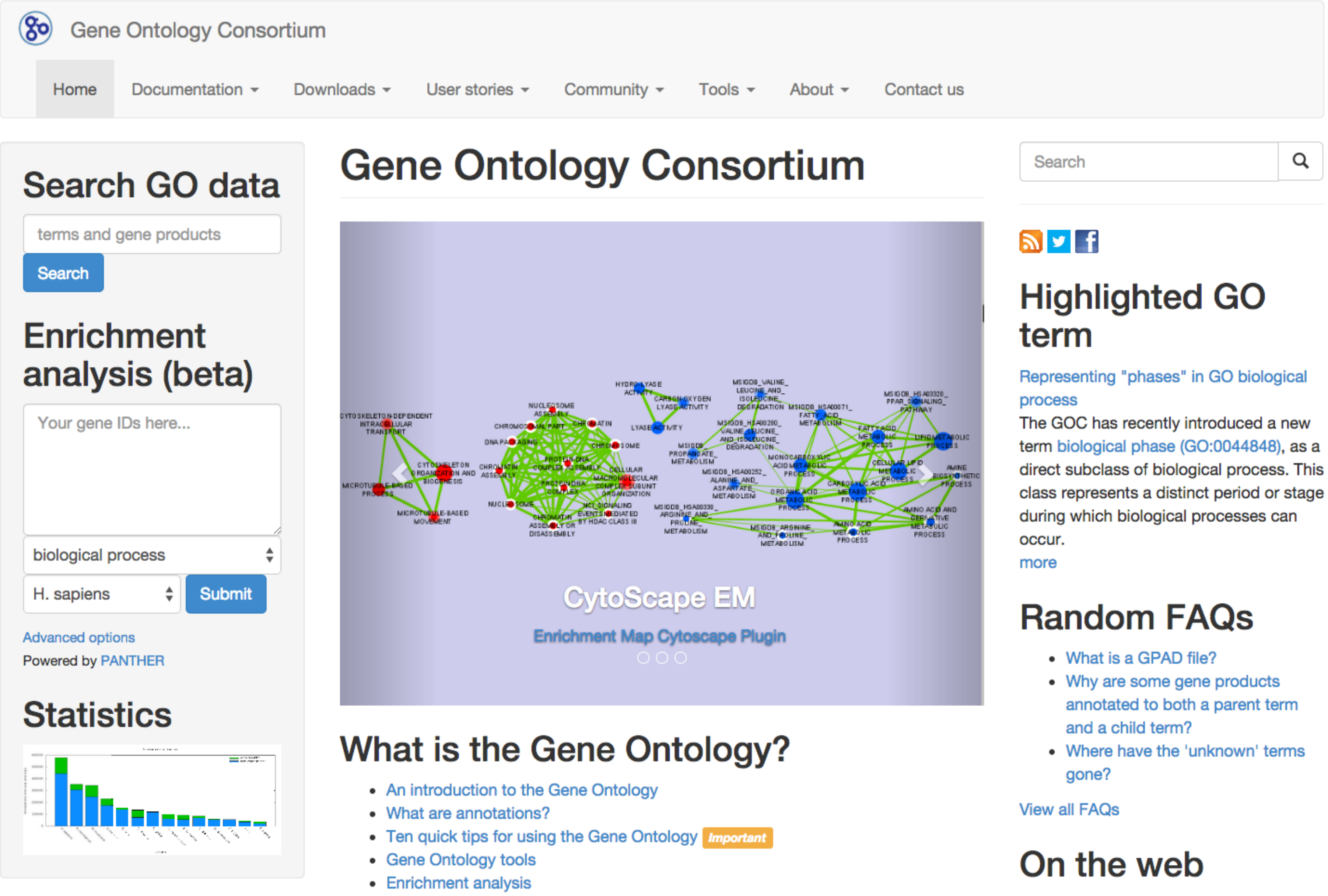
2. Use the list of SwissProt ids you get from the previous part and extract the corresponding sequences.

Modify the script in part a) to automatically download the sequence from the web and count the number or amino acids that compose each sequence.

Function & Computing

Can we transform functional annotation in computer readable information?

This is the main aim of the Gene Ontology (GO) Consortium



Search GO data

terms and gene products

Enrichment analysis (beta)

Your gene IDs here...

Advanced options

Powered by PANTHER

Statistics

Gene Ontology Consortium

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology *Important*
- Gene Ontology tools
- Enrichment analysis

Highlighted GO term

Representing "phases" in GO biological process

The GOC has recently introduced a new term biological phase (GO:0044848), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur.

more

Random FAQs

- What is a GPAD file?
- Why are some gene products annotated to both a parent term and a child term?
- Where have the 'unknown' terms gone?

View all FAQs

On the web

http://www.geneontology.org/

Gene Ontology

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data.

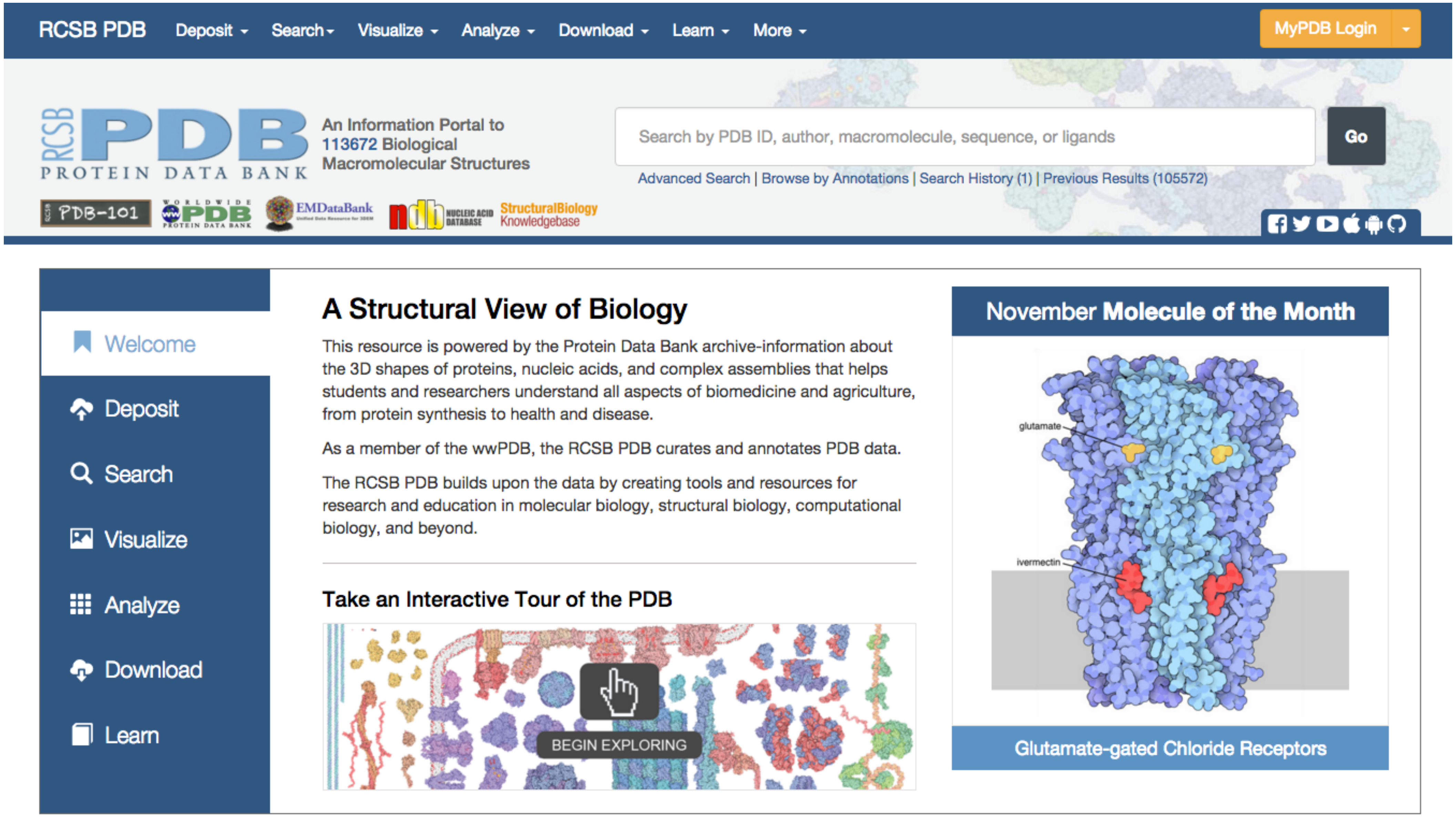
http://www.geneontology.org/

The ontology is represented by a direct acyclic graph covers three domains;

- cellular component, the parts of a cell or its extracellular environment (GO:0005575);
- molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis (GO:0003674)
- biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells,tissues, organs and organisms (GO:0008150).

The Protein Data Bank

The largest repository of macromolecular structures obtained mainly by X-ray crystallography and NMR



An Information Portal to 113672 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations | Search History (1) | Previous Results (105572)

A Structural View of Biology

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

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Take an Interactive Tour of the PDB

November Molecule of the Month

Glutamate-gated Chloride Receptors

http://www.pdb.org