Biological Data Sources and File Formats

Emidio Capriotti
http://biofold.org/emidio

Division of Informatics
Department of Pathology

Biomolecules Folding and Disease

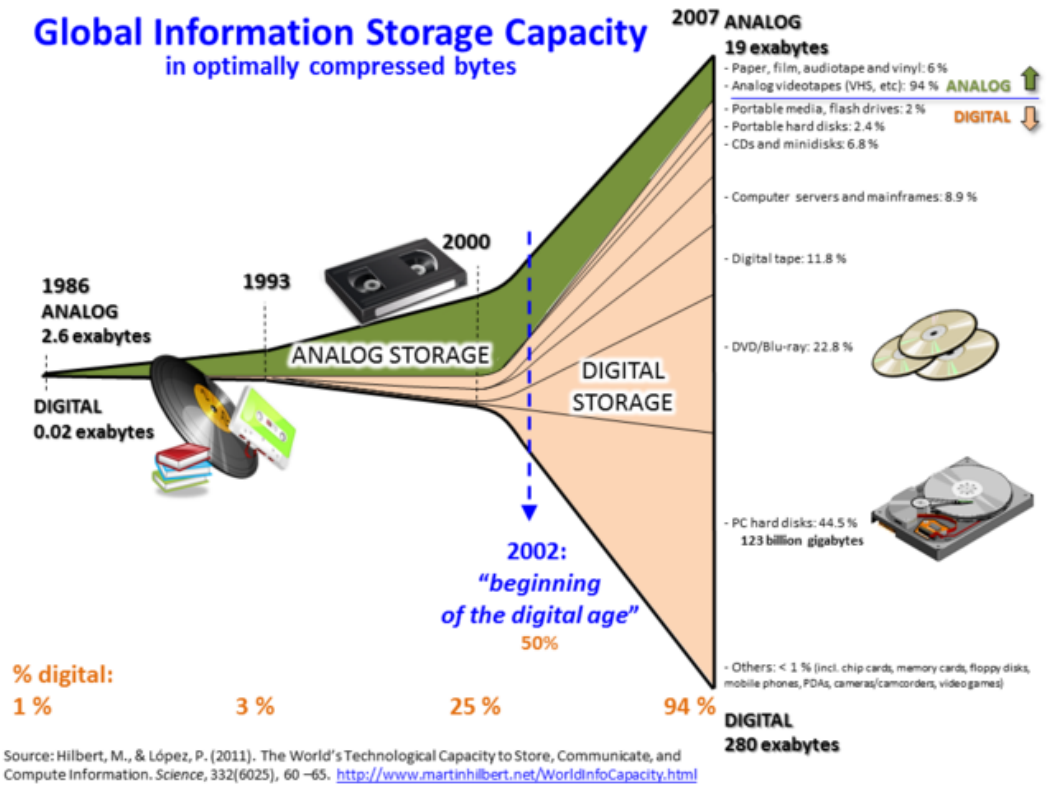
THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

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:

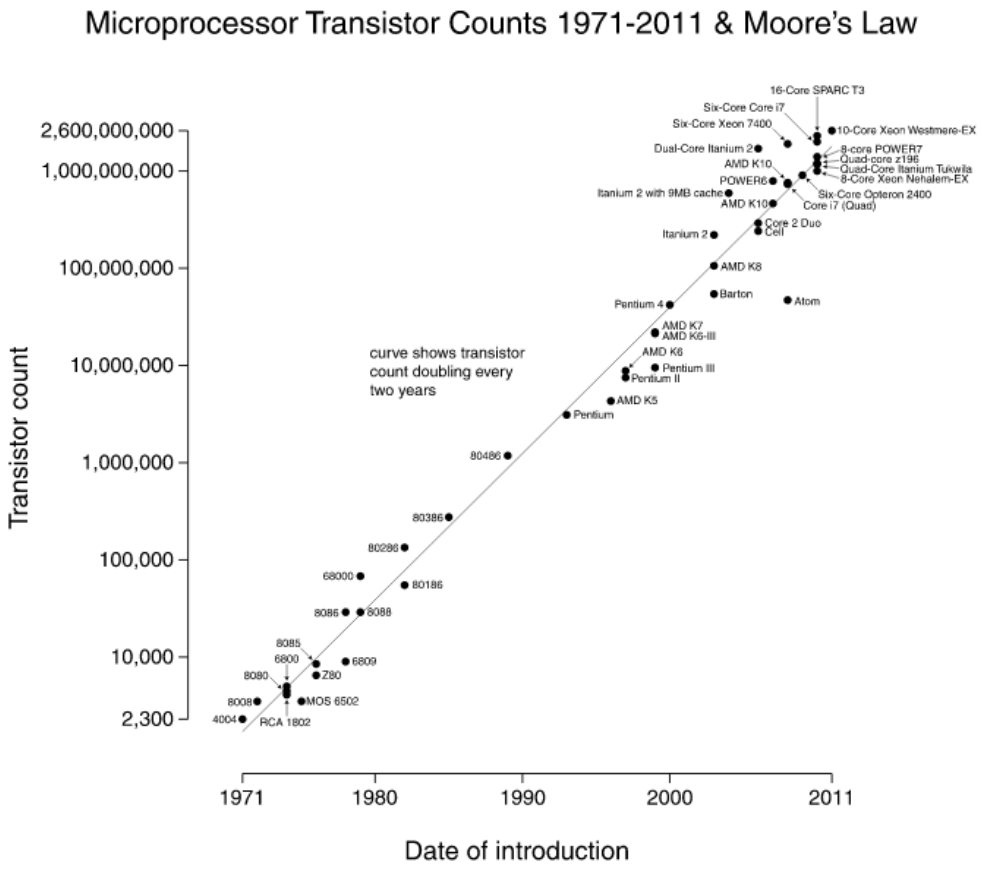
- analysis
- capture
- curation
- search
- sharing
- storage
- transfer
- visualization
- information privacy.



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

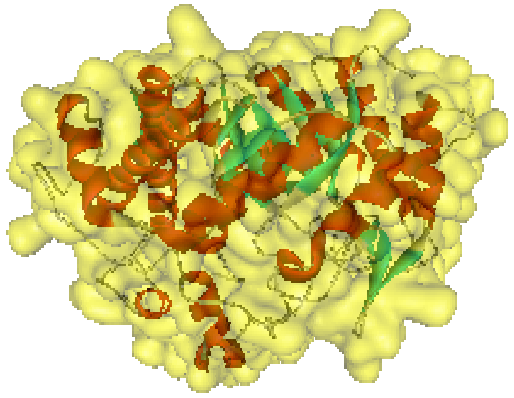
International consortiums such as HapMap, 1000Genomes and ENCODE are collecting large amount of data about the human genome.

The NCBI collects the complete genomic sequences of many organisms

- Archea: 195/473 species
- Bacteria: 3,421/31,028 species
- Eukariots: 20/1,924 species

Molecular biology data

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



GenBank:	179,295,769
UniRef90:	30,147,837
Swiss-Prot:	547,599
Protein Data Bank:	106,517
Protein:	98,954
Nucleic Acids:	2,749

Feb 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.

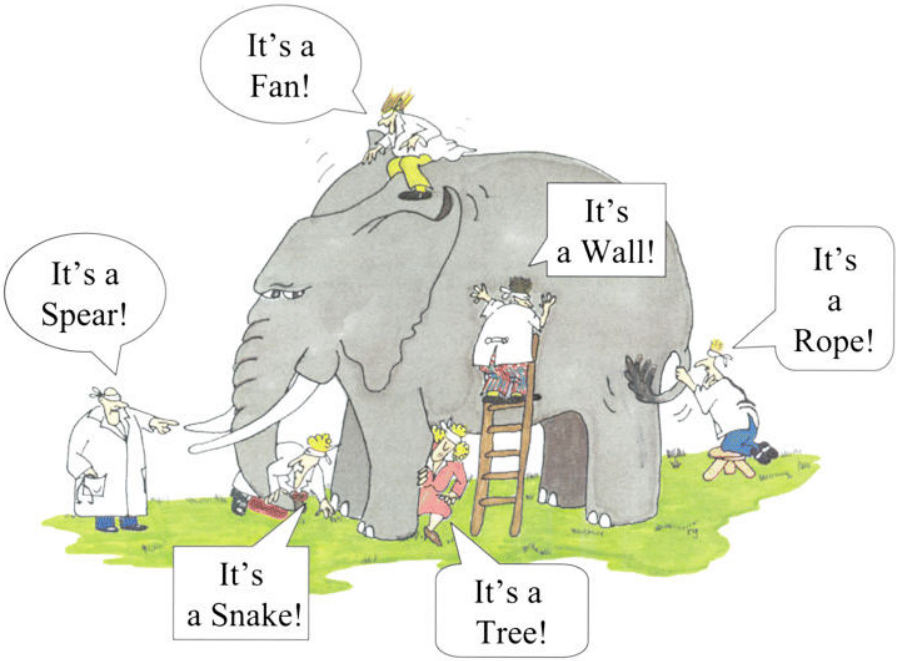
Computational biology involves the development and **application of data-analytical and theoretical methods**, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

from wikipedia

The elephant or the cave?

Computer sciences: building the big eye of the future

Scientists and the elephant



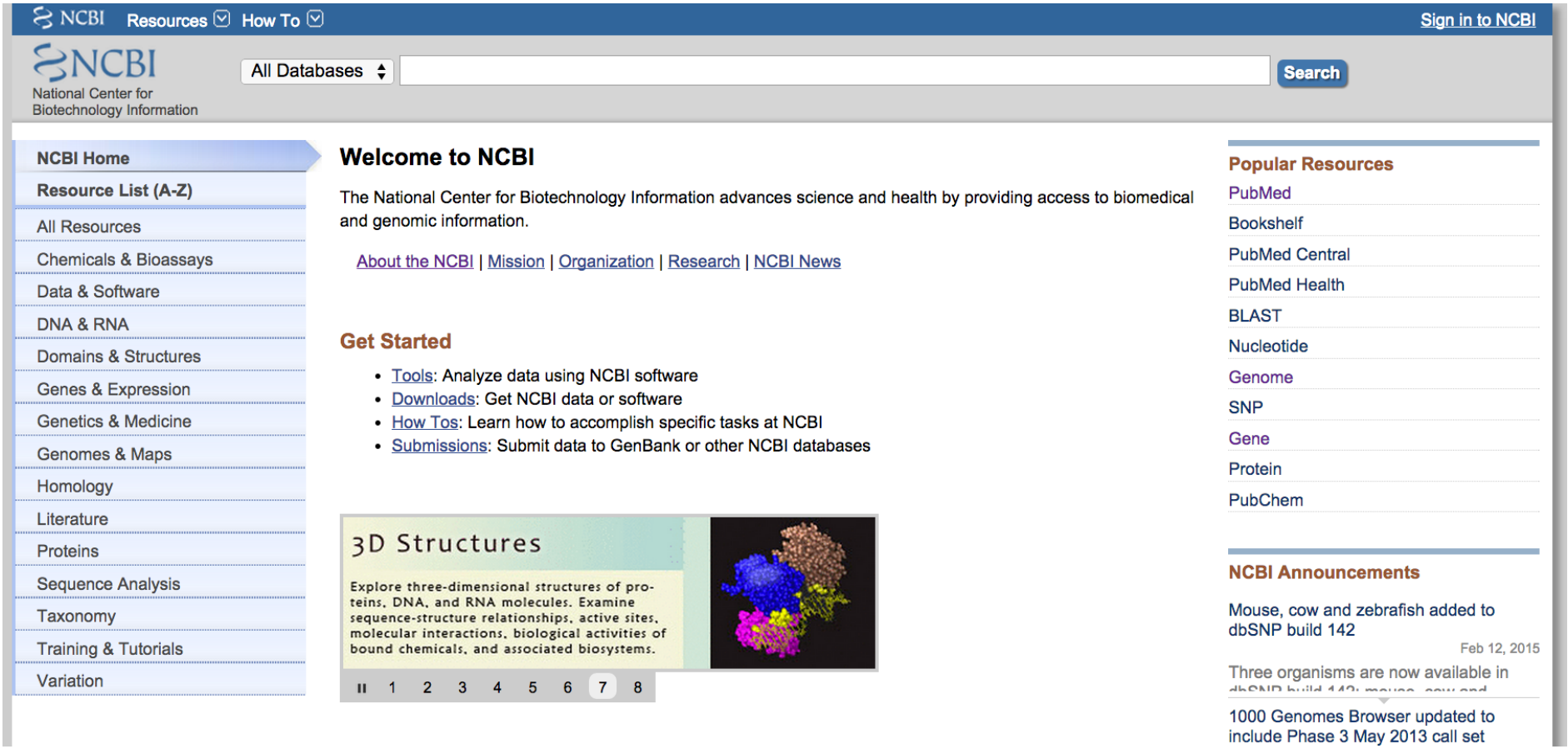
www.goldismoney2.com

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 …..



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:

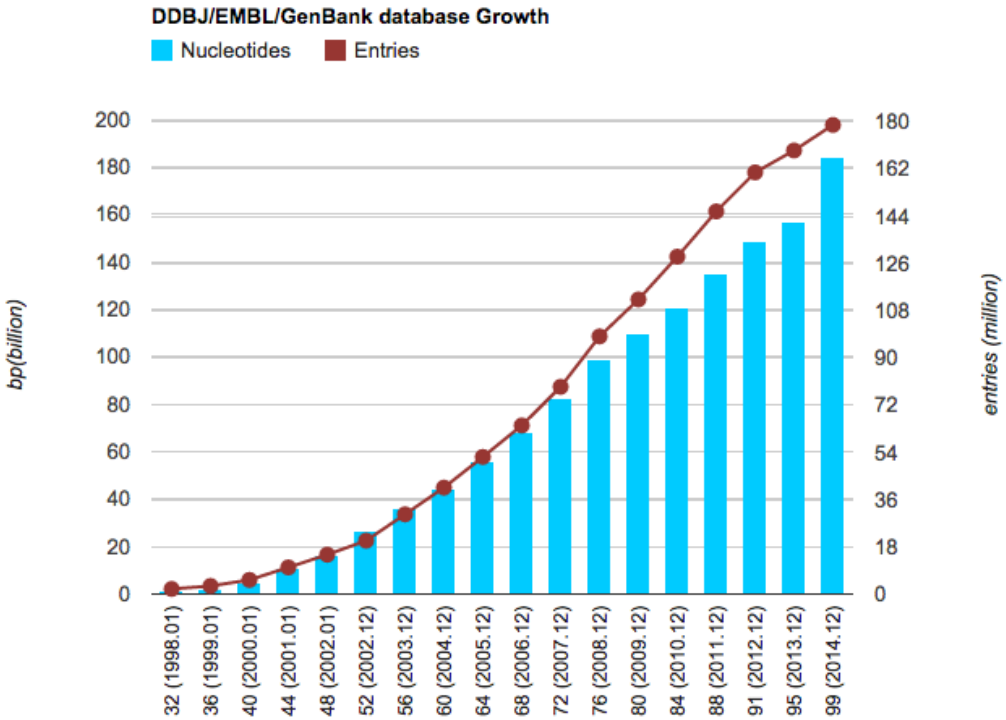
Homo sapiens tumor protein p53 (TP53), RefSeqGene (LRG_321) on chromosome 17

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.

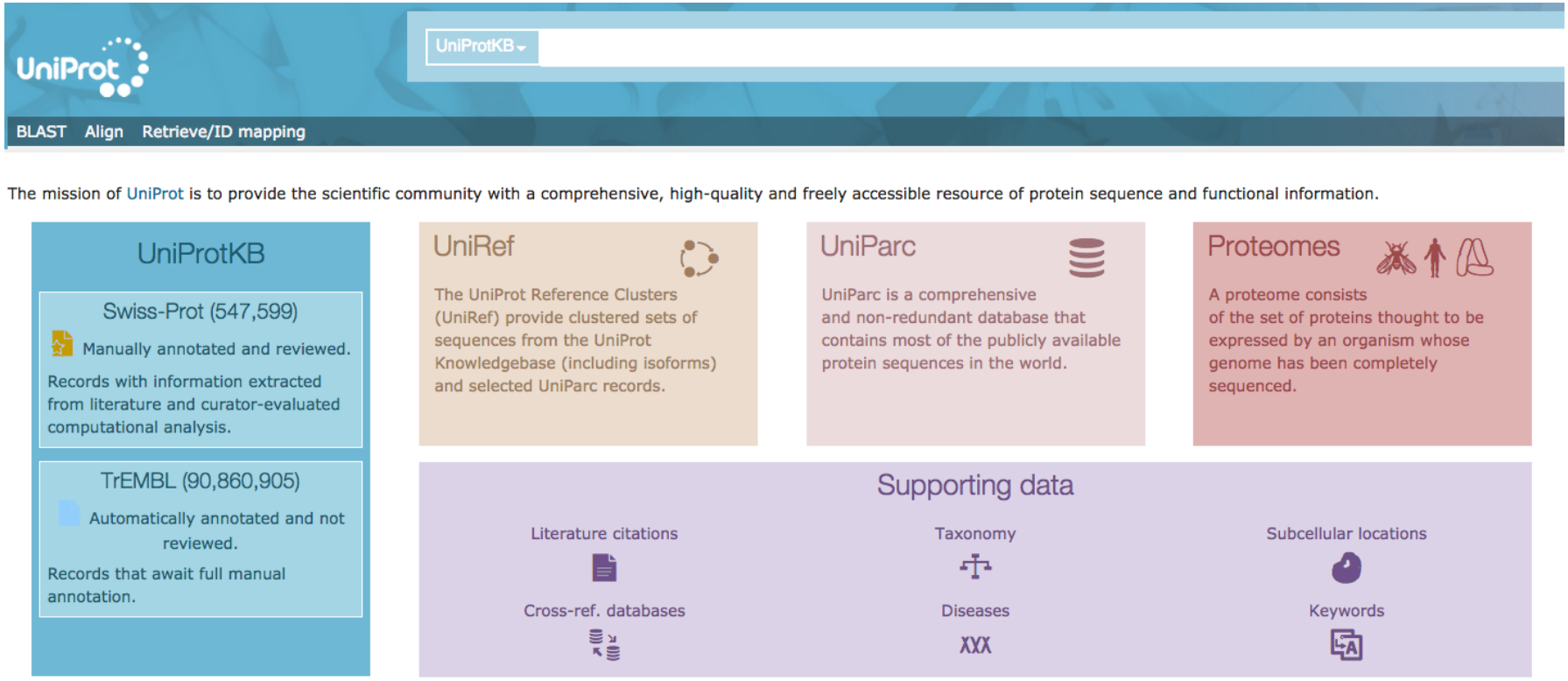
The annotation can be wither either automatic, using computational tools or manual, using results of experimental.

The automatic annotation is mainly based on homology search because

higher sequence similarity => higher the probability similarity in function

The UniProt

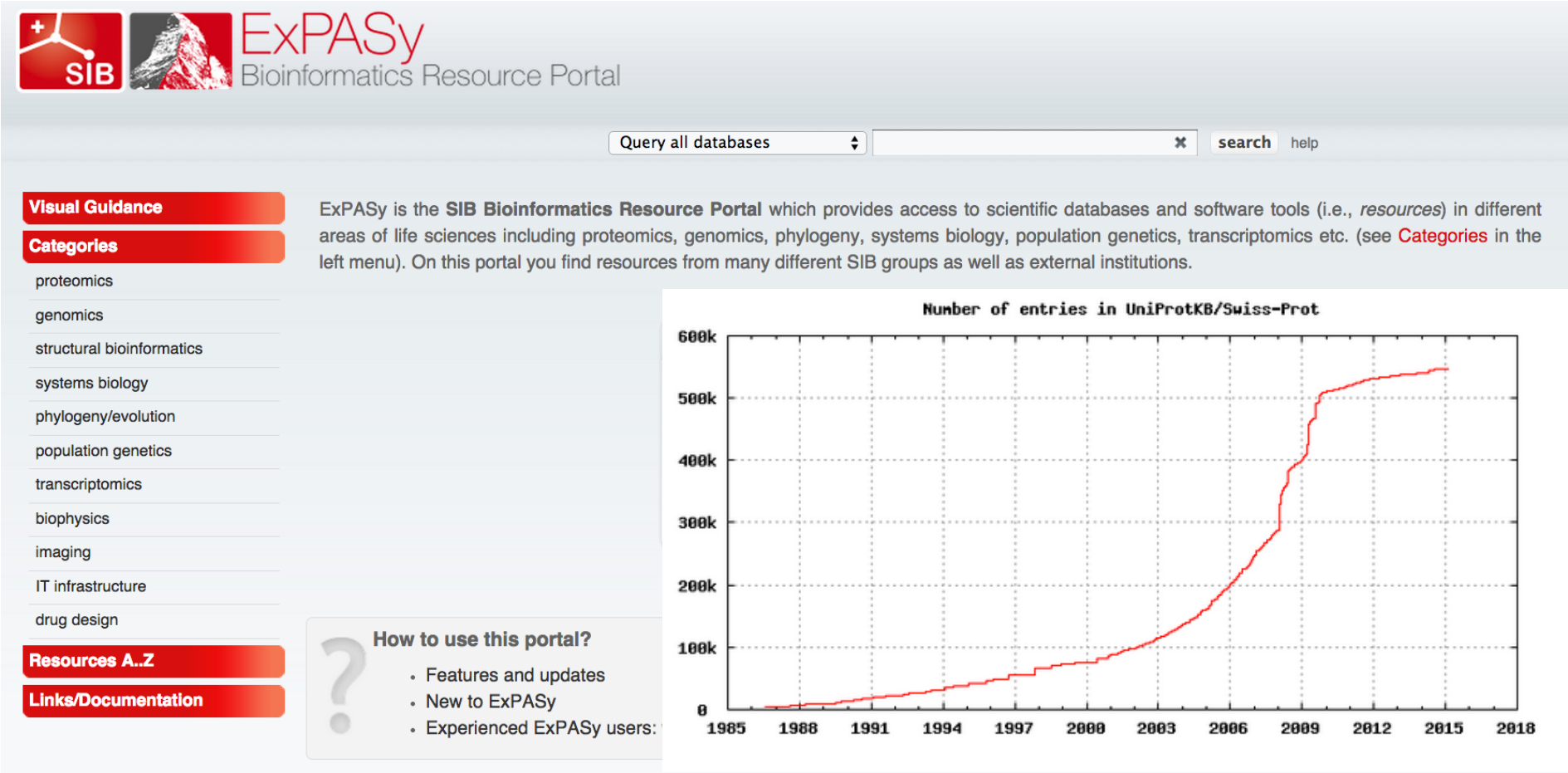
The European repository of molecular biology data. UniProtKB is composed by SwissProt and TrEMBL



http://www.uniprot.org/

The SwissProt

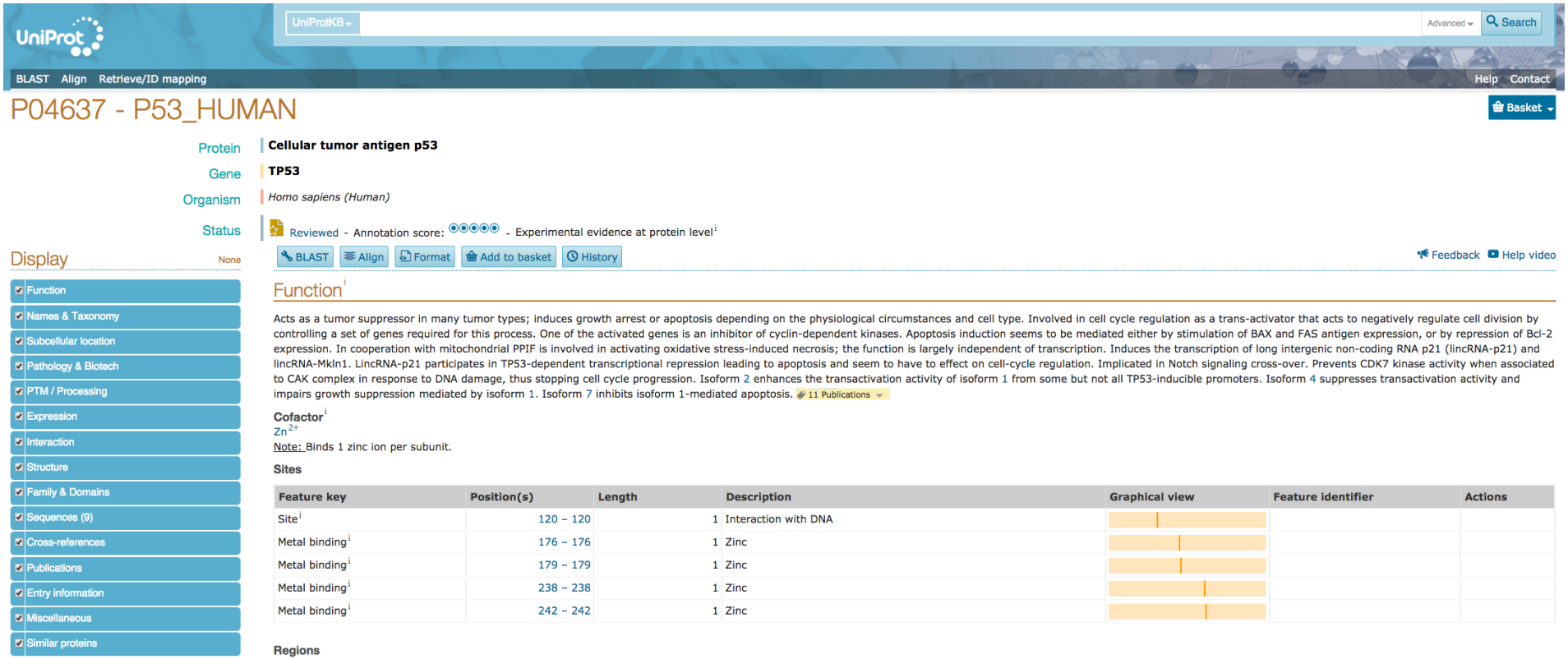
SwissProt contains all the proteins that have been manually annotated using information extracted from literature.



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



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/emidio/tmp/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.

Problem 2

Tyrosine kinase phosphorylation site (PS00007) is a common motif found in many protein sequences. The patter of the motif is defined with the following expression:

[RK]-x(2,3)-[DE]-x(2,3)-Y

Write a python script that scan all the sequences extract in the previous exercise to find if they contains the PS00007 motif and its possible locations

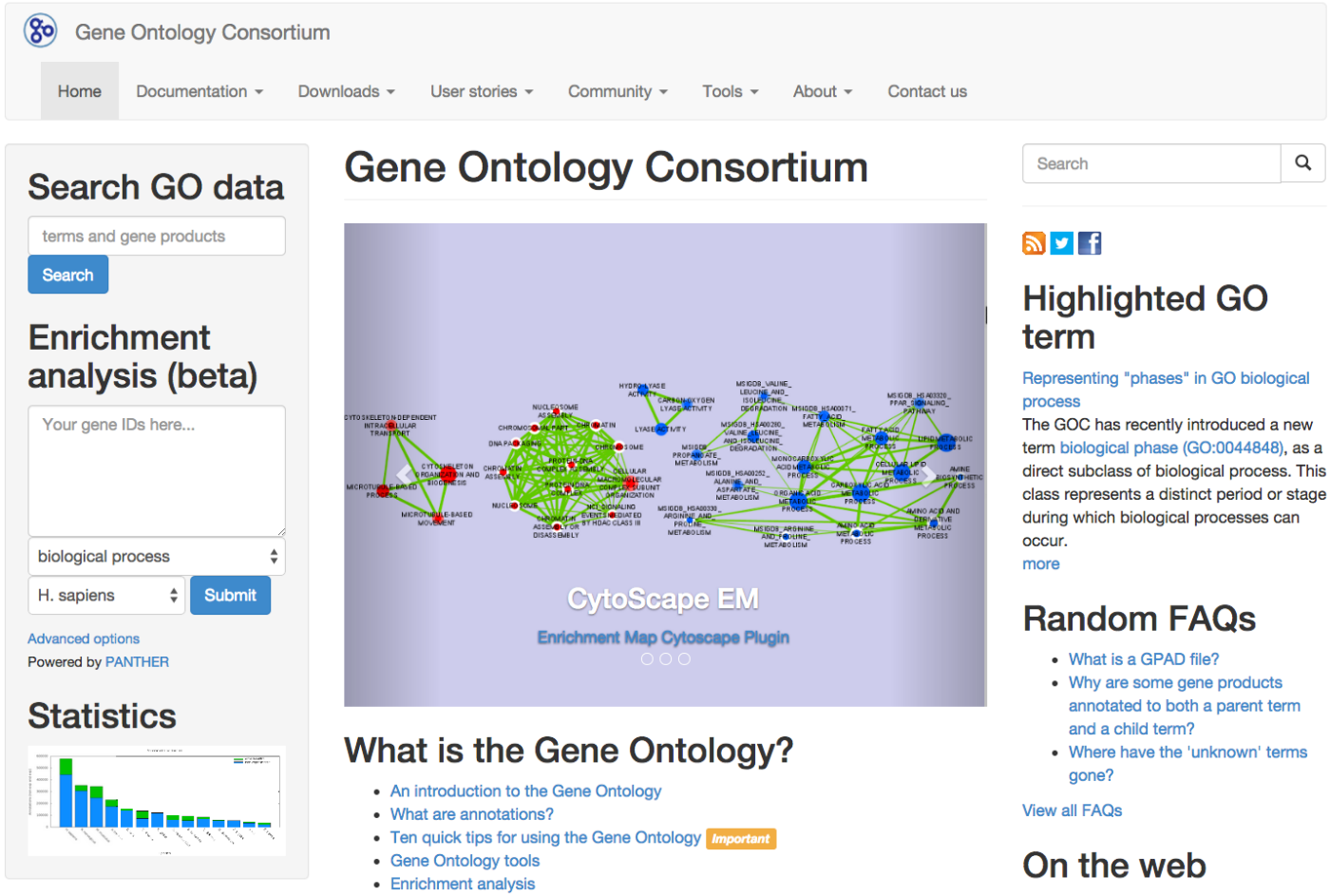
To solve this problem we use the **re module** in python and call the **finditer** method

More information about the Tyrosine phosphorylation site are available at: http://prosite.expasy.org/PS00007

Function & Computing

Can we transform functional annotation in computer readable information?

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



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.

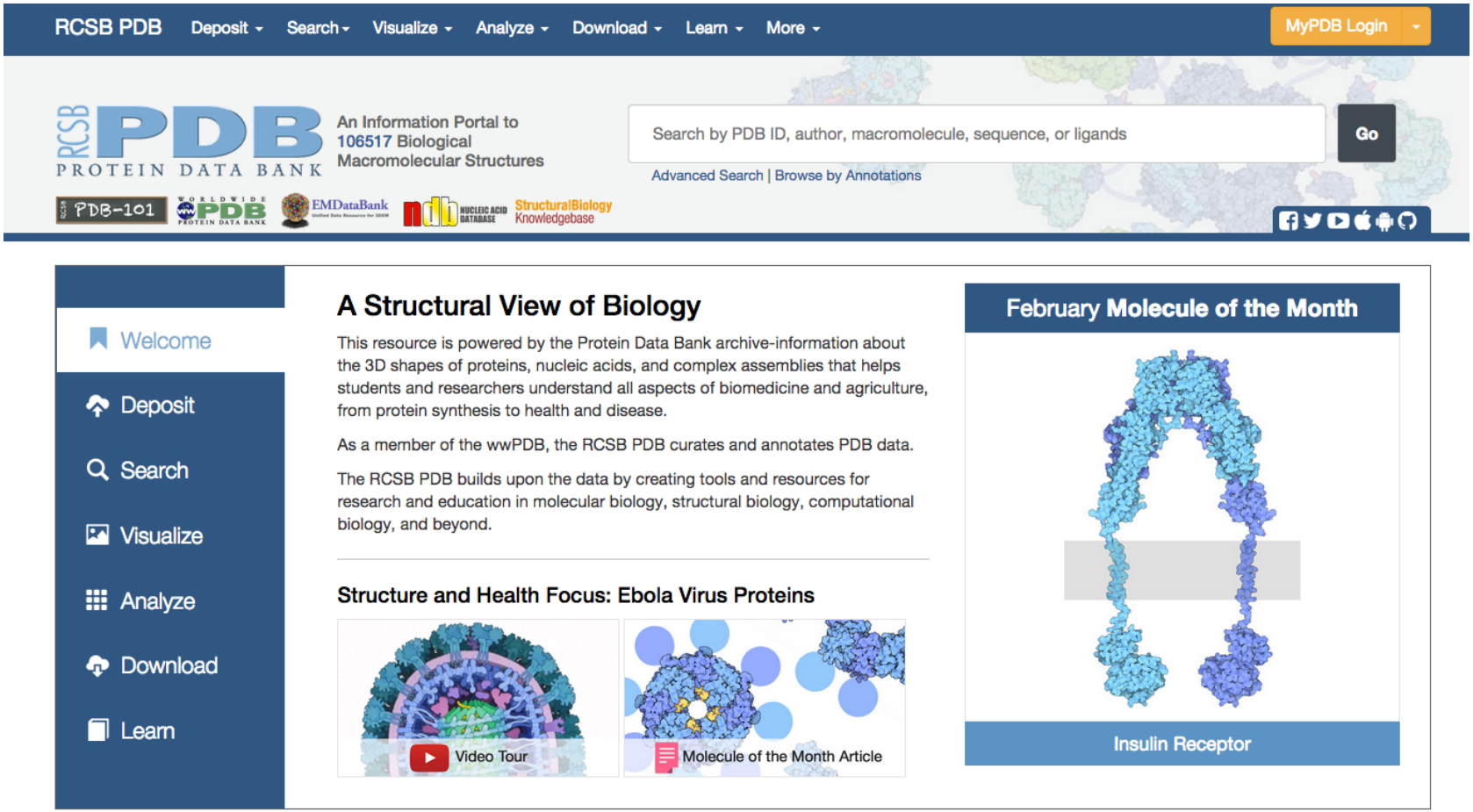
http://www.geneontology.org/

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



http://www.pdb.org