

Biological Data Sources and File Formats

CB2-201 – Computational Biology and Bioinformatics

February 17, 2015

Emidio Capriotti

<http://biofold.org/emidio>



**Biomolecules
Folding and
Disease**

Division of Informatics
Department of Pathology

UAB

THE UNIVERSITY OF
ALABAMA AT BIRMINGHAM

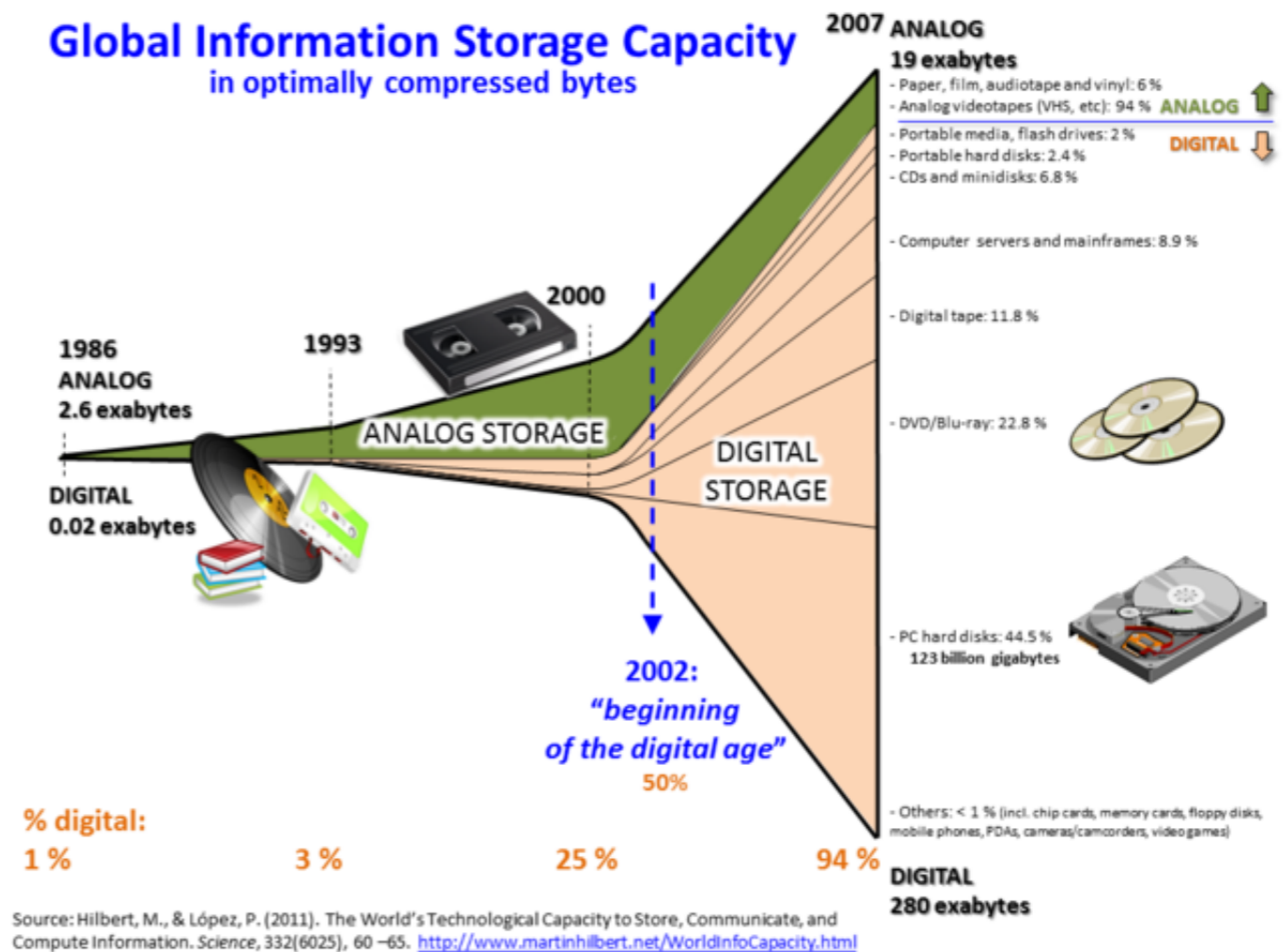


Big Data

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

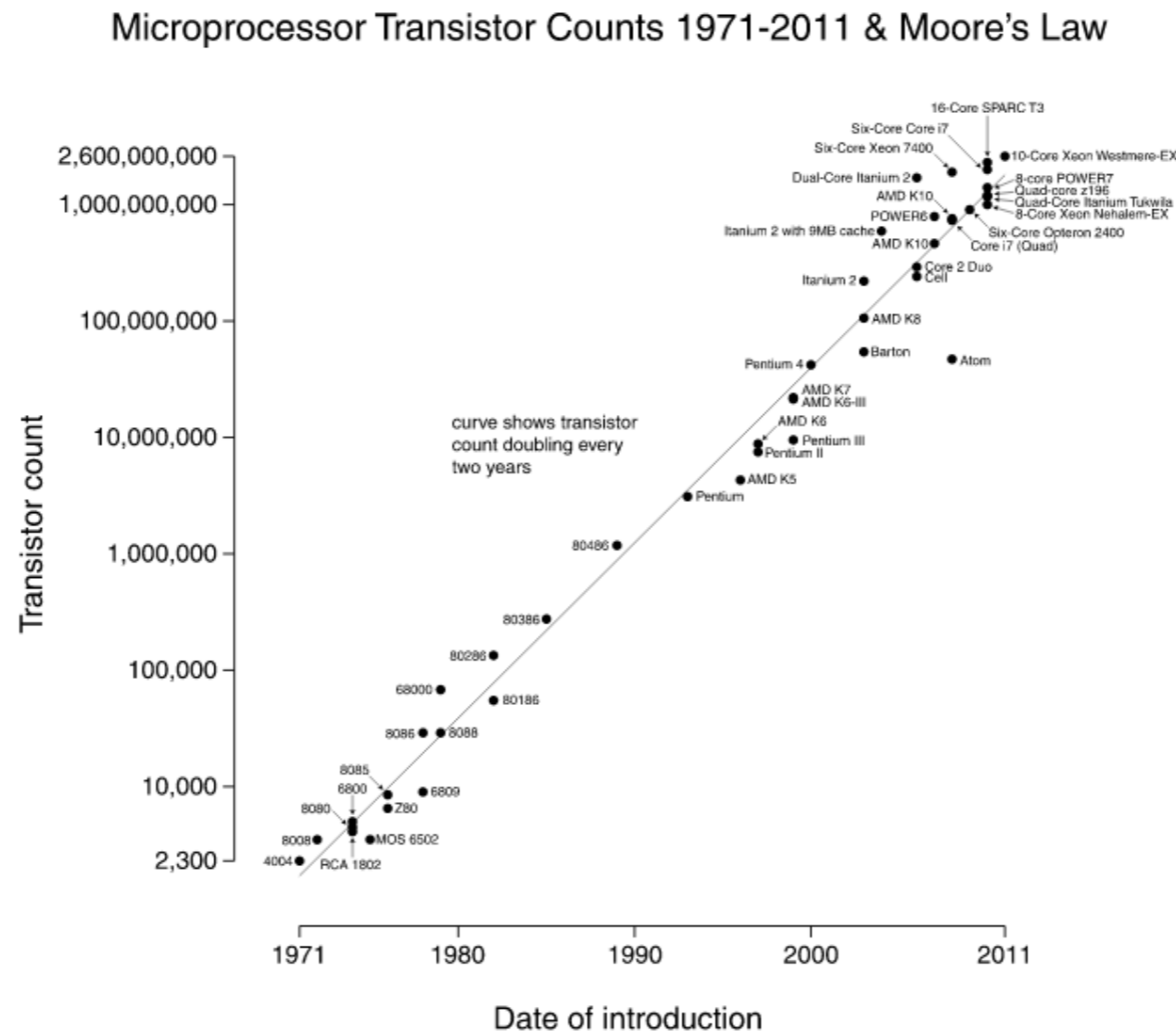
Main challenges include:

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



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

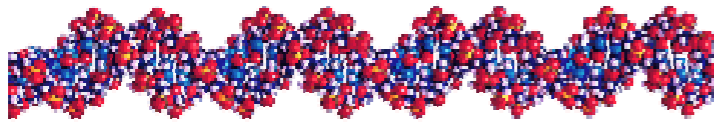
International HGS Consortium Nature 2004. PMID: 15496913

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

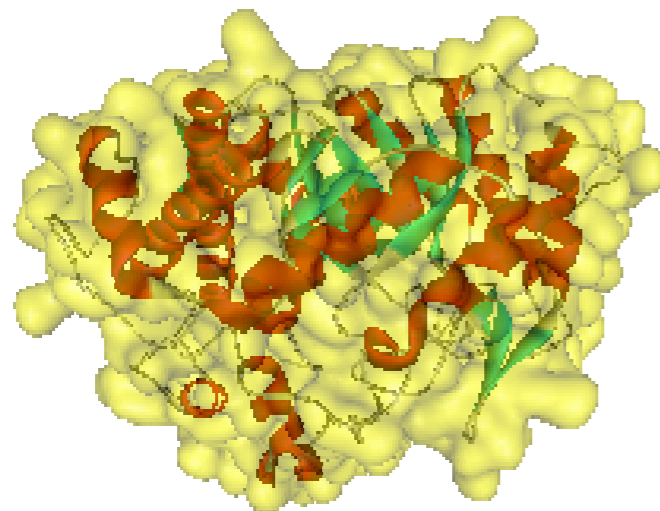


```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.  
MYSFPNSFRFGWSQAGFQSEMGTGSEDPNTDWYKWVHDPENMAAGLVSG  
DLPENGGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLRPQNFDE  
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH  
WPLPLWLHDPIRVRRGDFTPSGWLSTRTVYEFARFSAYIAWKFDDLVD  
YSTMNEPNVVGGLGYGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI  
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR  
GNEKIVRDDLKGRDLWIGVNYTRTVVKRTEKGYVSLGGYGHGCERNVS  
LAGLPTSDFGWEEFFPEGLYDVLTKYWNRHLYMYVTENGIADDADYQRPY  
YLVSHVYQVHRAINSGADVRYLHWSLADNYEWASGFSMRFLLKVDYNT  
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

Definition

Computational Biology and Bioinformatics: **the same focus but different priorities**

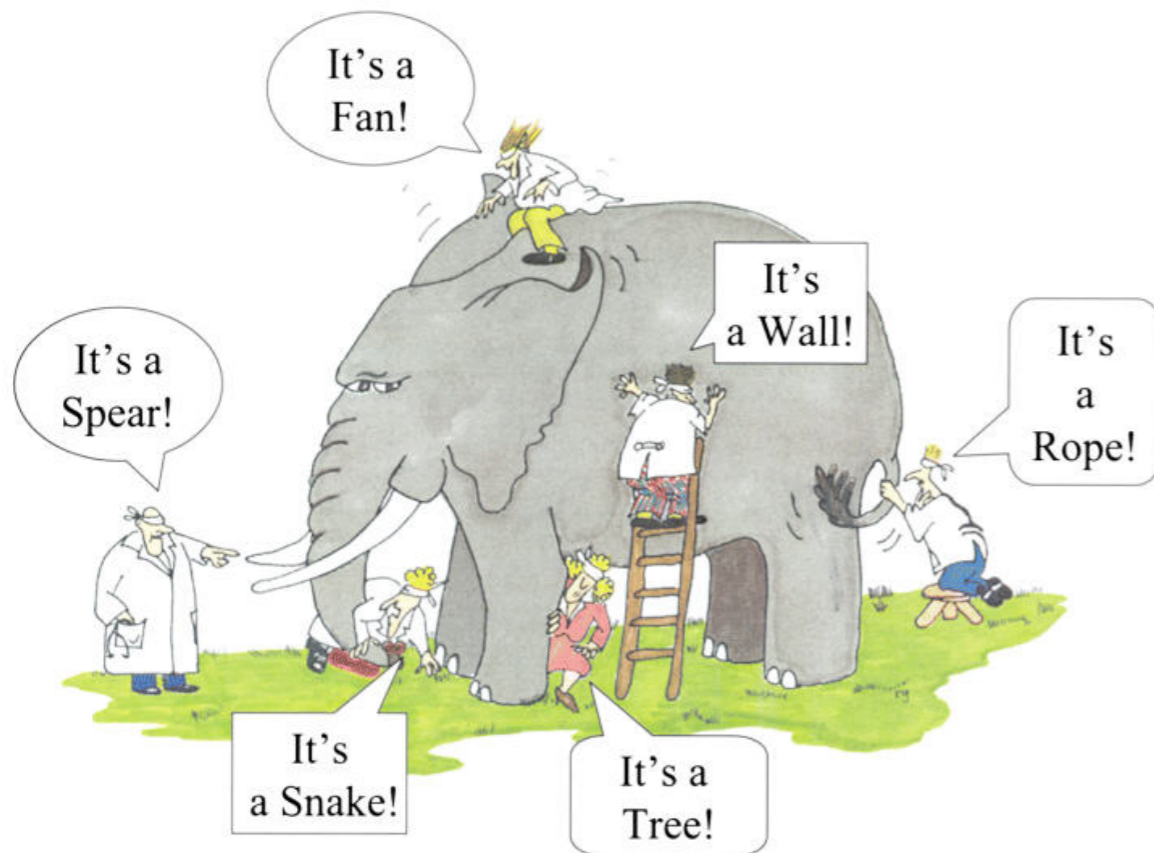
Bioinformatics is an interdisciplinary field that **develops methods and software tools for understanding biological data**. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering **to study and process biological data**.

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**.

The elephant or the cave?

Computer sciences: building the big eye of the future

Scientists and the elephant



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



The NCBI

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

The screenshot shows the NCBI homepage with a dark blue header containing the NCBI logo, navigation links for 'Resources' and 'How To', and a 'Sign in to NCBI' link. Below the header is a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. On the left is a vertical navigation menu with categories like 'NCBI Home', 'Resource List (A-Z)', and various biological topics. The main content area features a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How Tos, and Submissions, and a '3D Structures' section with a protein structure image and a description. On the right, there are sections for 'Popular Resources' (including PubMed, Bookshelf, etc.) and 'NCBI Announcements' (including news about dbSNP build 142 and the 1000 Genomes Browser).

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 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
CCCAGCTAATTTTTTTGTATTTTAAATAGAGATGGGGTTTCATCATGTTGGCCAGGCTGGTCTCGAACTCC
TGACCTCAGGTGATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAGGAGTCAGCCACCGCACCCA
.....
```

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

```
>P1;CRAB_ANAPL
ALPHA CRYSTALLIN B CHAIN (ALPHA (B)-CRYSTALLIN) .
MDITIHNP LIRRPLFSWLAPSRIFDQIFGEHLQESELLPASPSLS PFLMRSPIFRMP SWLETGLSEMRLEK
DKFSVNL DVKHFSPEELKV KVLGDMVEIHGKHEERQDEHGF IAREFN RKYRI PADVDPLTITSSLSLDGVL
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

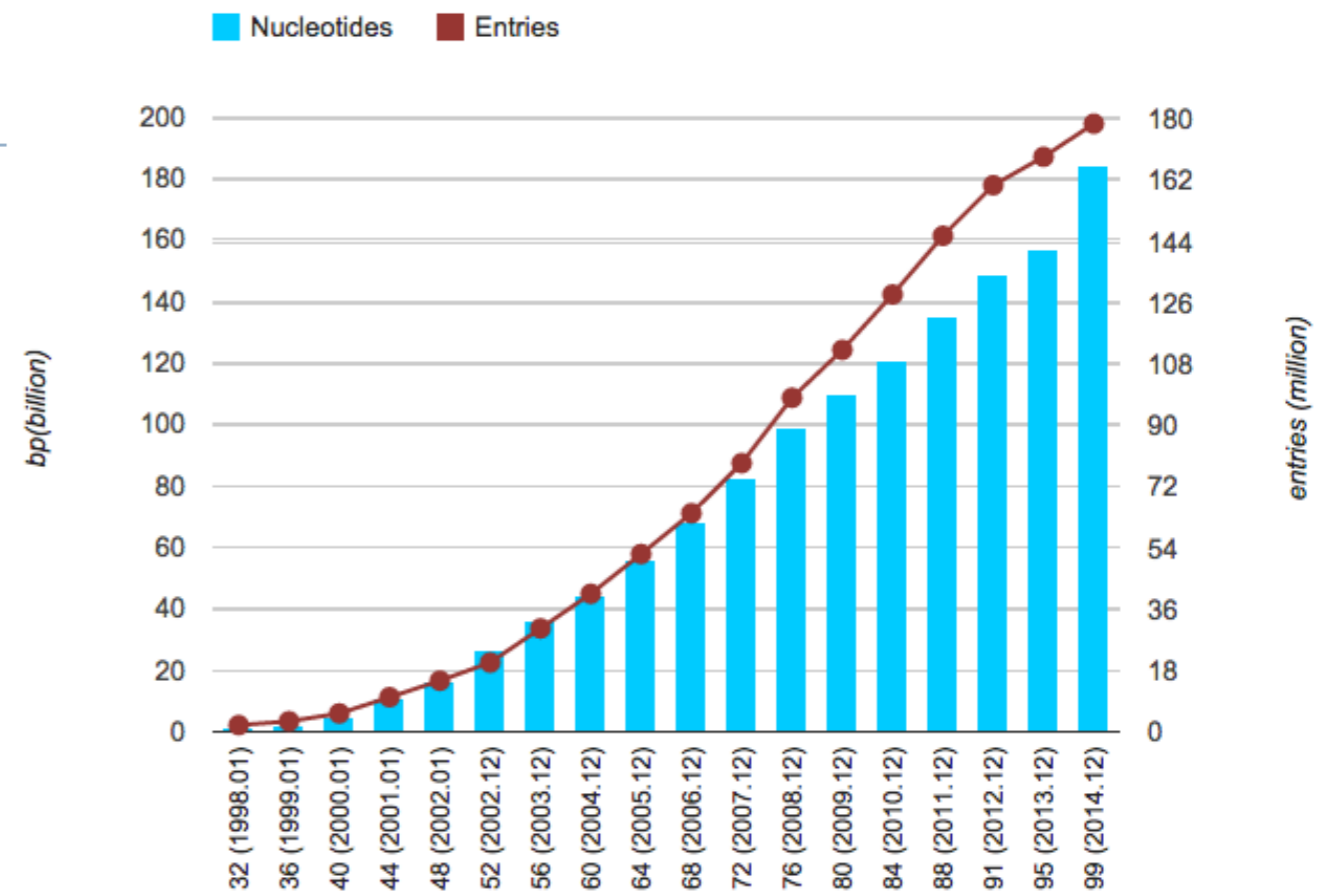
NCBI Reference Sequence: NG_017013.2

[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.

DDBJ/EMBL/GenBank database Growth



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.

<h3>UniProtKB</h3> <p>Swiss-Prot (547,599) Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis.</p> <p>TrEMBL (90,860,905) Automatically annotated and not reviewed. Records that await full manual annotation.</p>	<h3>UniRef</h3> <p>The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.</p>	<h3>UniParc</h3> <p>UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.</p>	<h3>Proteomes</h3> <p>A proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.</p>						
<h3>Supporting data</h3> <table><tr><td>Literature citations</td><td>Taxonomy</td><td>Subcellular locations</td></tr><tr><td>Cross-ref. databases</td><td>Diseases</td><td>Keywords</td></tr></table>				Literature citations	Taxonomy	Subcellular locations	Cross-ref. databases	Diseases	Keywords
Literature citations	Taxonomy	Subcellular locations							
Cross-ref. databases	Diseases	Keywords							

The SwissProt

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

Visual Guidance

Categories

- proteomics
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

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:

Number of entries in UniProtKB/Swiss-Prot

Year	Number of entries (k)
1985	0
1988	10
1991	20
1994	30
1997	40
2000	60
2003	100
2006	200
2009	400
2012	500
2015	550

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

UniProt
UniProtKB - [Search] Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

P04637 - P53_HUMAN

Basket

Protein Cellular tumor antigen p53
Gene TP53
Organism *Homo sapiens (Human)*
Status Reviewed - Annotation score: [5/5] - Experimental evidence at protein levelⁱ

BLAST Align Format Add to basket History Feedback Help video

Functionⁱ

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-Mkn1. 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](#)

Cofactorⁱ

Zn²⁺
Note: Binds 1 zinc ion per subunit.

Sites

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

Regions

Display None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences (9)
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similar proteins

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 Vogelstein 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 from
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 of amino acids that compose each sequence.

Problem 2

Tyrosine kinase phosphorylation site (PS00007) is a common motif found in many protein sequences. The pattern 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?

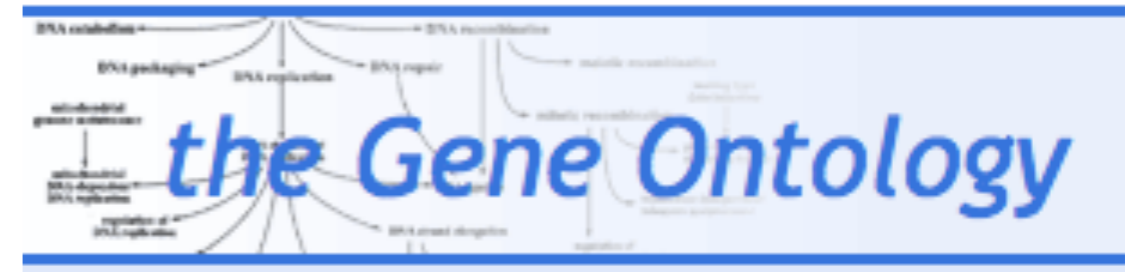
This is the main aim of the **Gene Ontology (GO) Consortium**

The screenshot displays the Gene Ontology Consortium website. At the top, there is a navigation menu with links for Home, Documentation, Downloads, User stories, Community, Tools, About, and Contact us. Below the navigation is a search bar with the text "Search" and a magnifying glass icon. The main content area is divided into several sections:

- Search GO data:** A search box containing "terms and gene products" and a blue "Search" button.
- Enrichment analysis (beta):** A section with a text input field "Your gene IDs here...", a dropdown menu set to "biological process", and a dropdown menu set to "H. sapiens". A blue "Submit" button is located to the right of the "H. sapiens" dropdown. Below this, there are links for "Advanced options" and "Powered by PANTHER".
- Statistics:** A section featuring a bar chart with a legend and a title "Statistics".
- Gene Ontology Consortium:** A large section with a central image of a network graph titled "CytoScape EM Enrichment Map Cytoscape Plugin". The graph shows various GO terms connected by lines, with some terms highlighted in red and blue. Below the graph, there is a heading "What is the Gene Ontology?" followed by a list of bullet points: "An introduction to the Gene Ontology", "What are annotations?", "Ten quick tips for using the Gene Ontology" (with an "Important" tag), "Gene Ontology tools", and "Enrichment analysis".
- Highlighted GO term:** A section with social media icons (RSS, Twitter, Facebook) and a heading "Highlighted GO term". Below the heading, there is a paragraph of text: "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." Below the paragraph is a blue link "more".
- Random FAQs:** A section with a heading "Random FAQs" and a list of three bullet points: "What is a GPAD file?", "Why are some gene products annotated to both a parent term and a child term?", and "Where have the 'unknown' terms gone?". Below the list is a blue link "View all FAQs".
- On the web:** A section with a heading "On the web".

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

The screenshot shows the top navigation bar of the RCSB PDB website. It includes a dark blue header with white text for navigation: "RCSB PDB", "Deposit", "Search", "Visualize", "Analyze", "Download", "Learn", and "More". On the right side of the header is a yellow "MyPDB Login" button. Below the header is a light blue banner with the "RCSB PDB PROTEIN DATA BANK" logo on the left. To the right of the logo is the text "An Information Portal to 106517 Biological Macromolecular Structures". A search bar is positioned in the center of the banner, containing the text "Search by PDB ID, author, macromolecule, sequence, or ligands" and a "Go" button. Below the search bar are links for "Advanced Search" and "Browse by Annotations". At the bottom of the banner are several logos: "PDB-101", "WORLDWIDE PDB PROTEIN DATA BANK", "EMDataBank", "NUCLEIC ACID DATABASE", and "StructuralBiology Knowledgebase". On the far right of the banner are social media icons for Facebook, Twitter, YouTube, Apple, Android, and RSS.

The screenshot shows the main content area of the RCSB PDB website. On the left is a dark blue sidebar with white text and icons for navigation: "Welcome", "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn". The main content area has a white background. At the top left of the main content is the heading "A Structural View of Biology". Below this heading is a paragraph: "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." Below this paragraph is another paragraph: "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." Below this text is a section titled "Structure and Health Focus: Ebola Virus Proteins". This section contains two images: a "Video Tour" showing a 3D model of a protein structure and a "Molecule of the Month Article" showing a 3D model of a protein structure. On the right side of the main content area is a section titled "February Molecule of the Month". This section features a large 3D model of the Insulin Receptor protein structure, which is colored in shades of blue and purple. Below the model is the text "Insulin Receptor".

<http://www.pdb.org>