

# Biological Data Sources and File Formats

**Bioinformatics fo Systems and Synthetic Biology**

**Emidio Capriotti**

<http://biofold.org/emidio>



**Biomolecules  
Folding and  
Disease**

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna

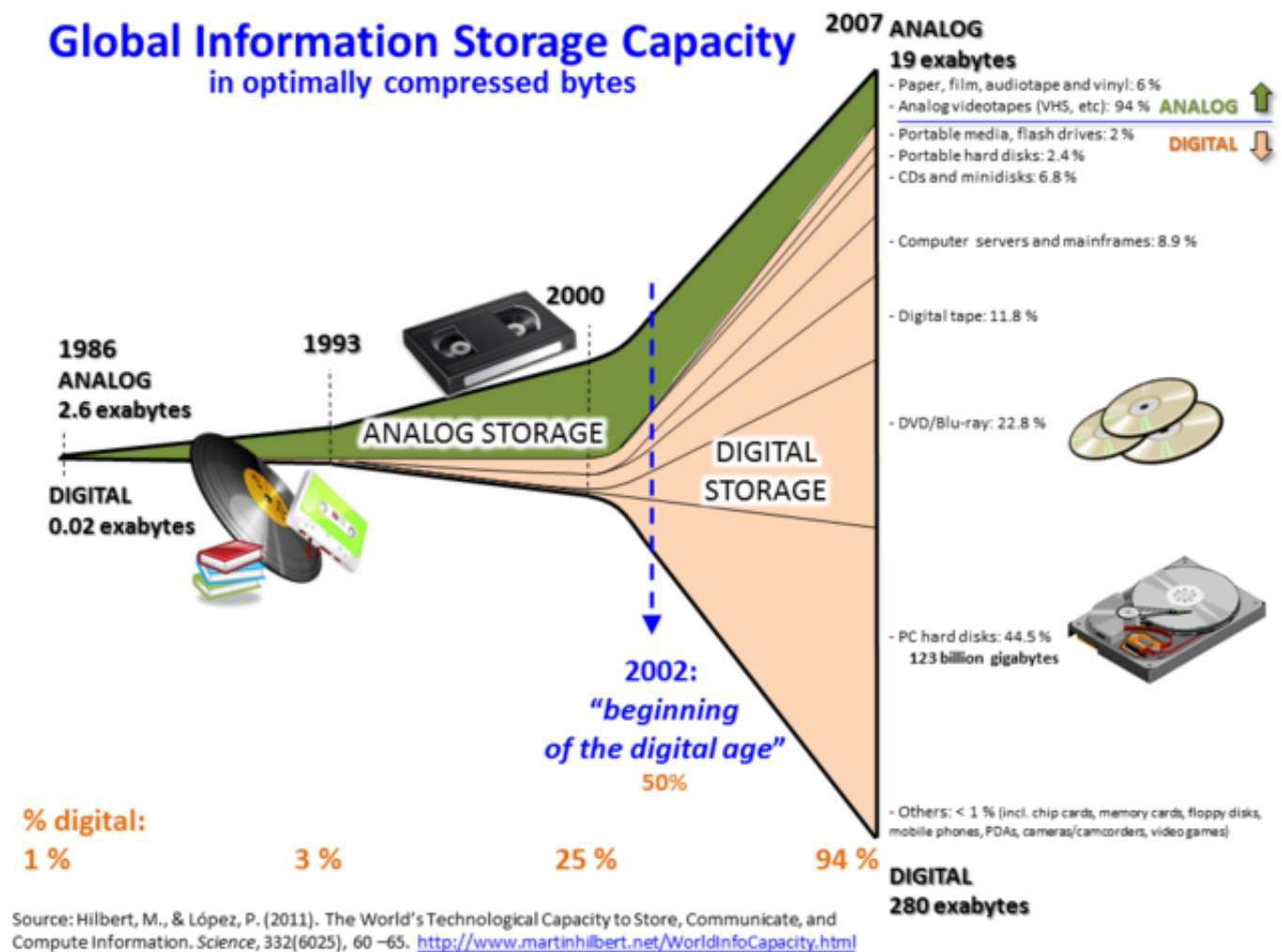


# Big Data

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

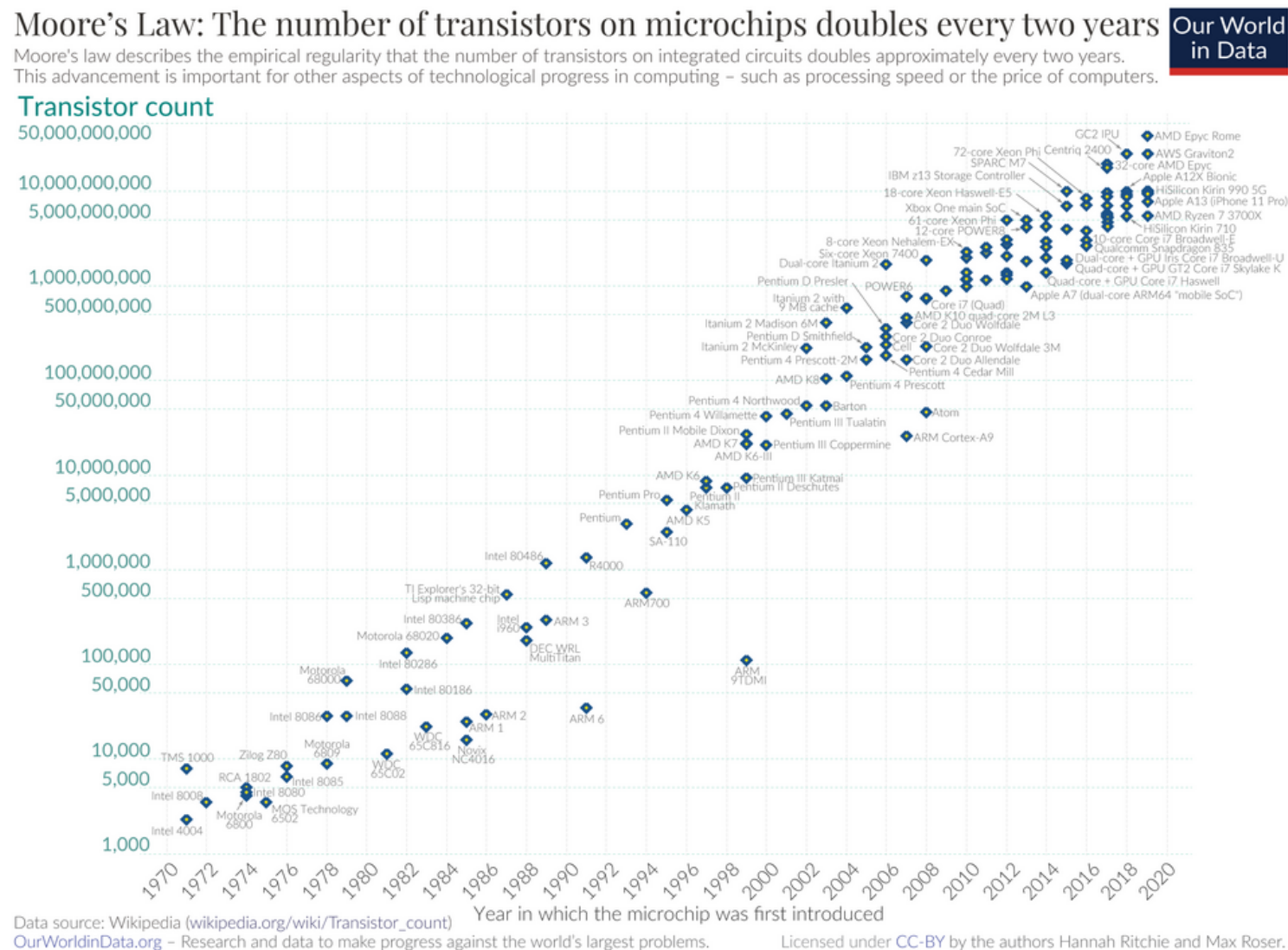
Main challenges include:

- 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.



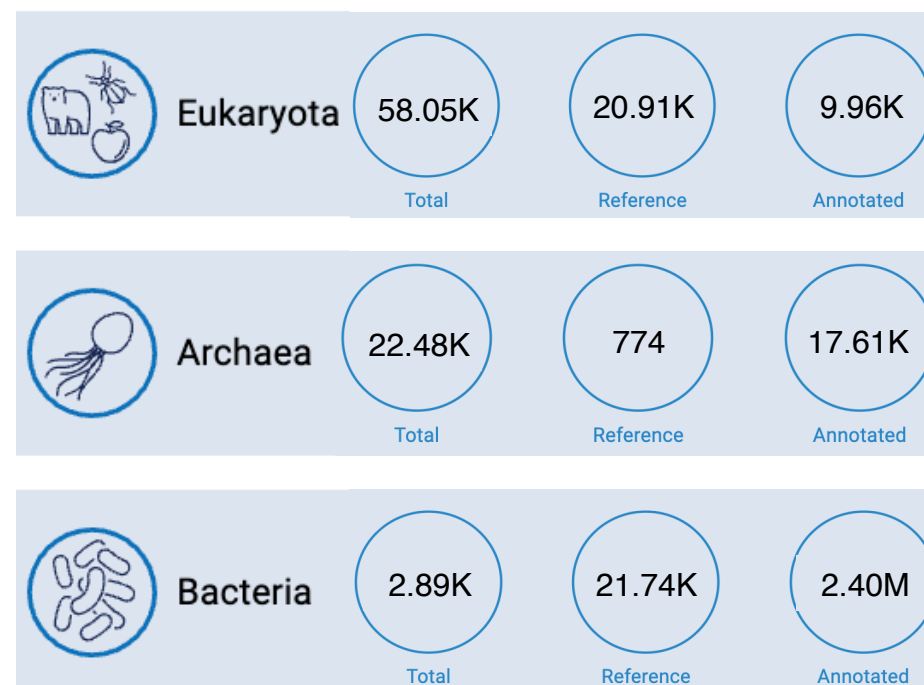
# Big Data in biology

The **complete human genome in the 2004** was released in 2004

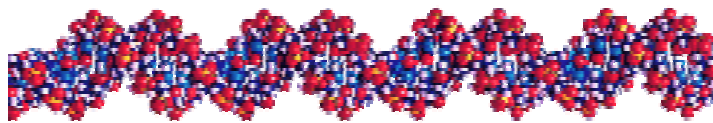
International HGS Consortium Nature 2004. PMID: 15496913

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**



# Molecular biology data



```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.  
MYSFPNSFRFGWSQAGFQSEMGTGSEDPNTDWYKWVHDPENMAAGLVSG  
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE  
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH  
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDLVDLVE  
YSTMNEPNVVGGLGYVGKSGFPFPGYLSFELSRRHMYNIIQAHARAYDGI  
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR  
GNEKIVRDDLKGRLDWIGVNYTTRTVVKRTEKGYVSLGGYGHGCERNVS  
LAGLPTSDFGWEEFFPEGLYDVLTKYWNRHYLYMYVTENGIADDADYQRPY  
YLVSHVYQVHRAINSGADVRYLHWSLADNYEWASGFSMRFGLLKVDYNT  
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

GenBank:

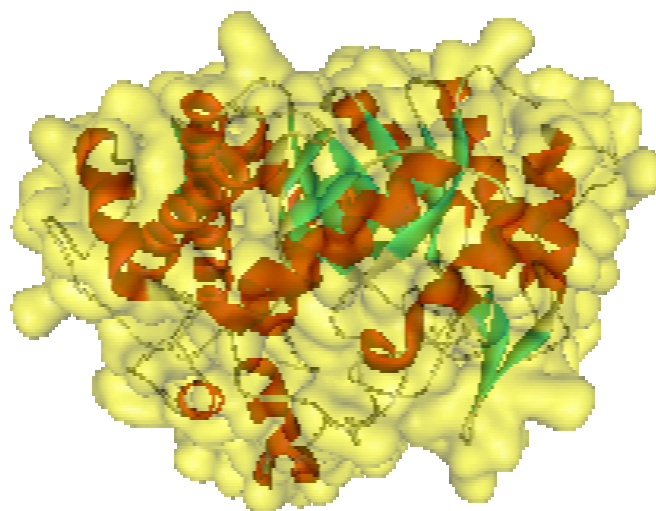
259,677,058

UniRef90:

184,146,434

Swiss-Prot:

573,661



Protein Data Bank:

248,329

Protein:

243,296

Nucleic Acids:

22,200

# The NCBI

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

The screenshot shows the NCBI National Library of Medicine homepage. At the top, there is a dark blue header with the NIH logo and the text "National Library of Medicine" and "National Center for Biotechnology Information". On the right of the header, there is a user profile icon and the email address "emidio.capriotti@...". Below the header, there is a search bar with a dropdown menu set to "All Databases" and a "Search" button. On the left side, there is a vertical navigation menu with the following items: "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", and "Variation". The main content area is titled "Welcome to NCBI" and contains the text: "The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information." Below this text, there are six main sections arranged in a 2x3 grid: "Submit" (Deposit data or manuscripts into NCBI databases), "Download" (Transfer NCBI data to your computer), "Learn" (Find help documents, attend a class or watch a tutorial), "Develop" (Use NCBI APIs and code libraries to build applications), "Analyze" (Identify an NCBI tool for your data analysis task), and "Research" (Explore NCBI research and collaborative projects). Each section has a corresponding icon. On the right side, there is a "Popular Resources" section with links to PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below this, there is an "NCBI News & Blog" section with two news items: "BankIt Submitters: Upcoming Changes to How You Submit to GenBank" dated 27 Jan 2026, and "GenBank Now Supports EGAPx-Based Annotation" dated 14 Jan 2026.

NIH National Library of Medicine  
National Center for Biotechnology Information

emidio.capriotti@...

All Databases Search

**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](#) | [NCBI News & Blog](#)

**Submit**  
Deposit data or manuscripts into NCBI databases

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class or watch a tutorial

**Develop**  
Use NCBI APIs and code libraries to build applications

**Analyze**  
Identify an NCBI tool for your data analysis task

**Research**  
Explore NCBI research and collaborative projects

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

**NCBI News & Blog**  
BankIt Submitters: Upcoming Changes to How You Submit to GenBank  
27 Jan 2026  
Are you a GenBank submitter? Do you use BankIt or the GenBank app in the  
GenBank Now Supports EGAPx-Based Annotation  
14 Jan 2026  
With the latest release of EGAPx, we're excited to announce

# 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) .
MDITIHNPILIRPLFSWLAPSRIFDQIFGEHLQESELLPASPSLSPFLMRSPIFRMPSWLETGLSEMRLEK
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

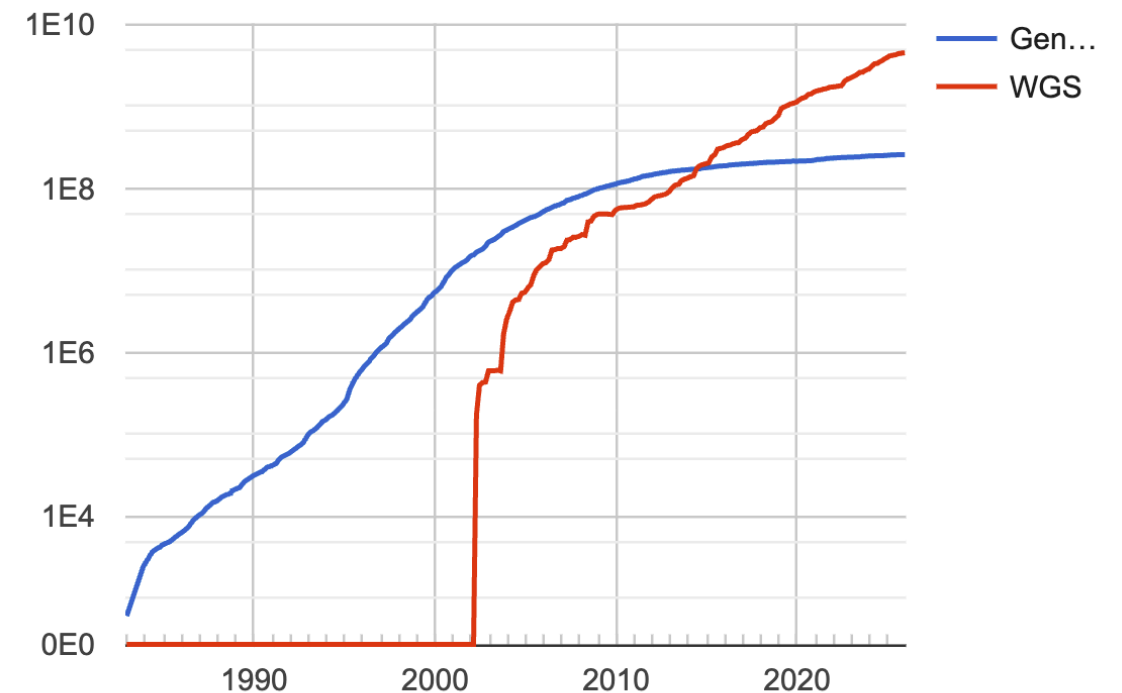
NCBI Reference Sequence: NG\_017013.2

[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.

### Sequences



# 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

## Find your protein

UniProtKB ▾

Advanced | List

Search

Examples: Insulin, APP, Human, P05067, organism\_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)”

⚠ Our Proteomes and UniProtKB/TrEMBL resources are undergoing a significant transition. Please read our [help page](#), view [affected entries and proteomes](#) [↗](#), or [contact us](#) with any questions.

### Proteins

UniProt Knowledgebase

Reviewed  
(Swiss-Prot)  
573,661

Unreviewed  
(TrEMBL)  
199,006,239

### Species

Proteomes

Protein sets for species with sequenced genomes from across the tree of life

### Protein Clusters

UniRef

Clusters of protein sequences at 100%, 90% & 50% identity

### Sequence

archive

UniParc

Non-redundant archive of publicly available protein sequences seen across different databases

# The SwissProt

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

[Home](#)[About](#)[Help](#)[ExpasyGPT](#)[SIB News](#)[Contact](#)

Posted 22 January 2026 - **Nextstrain** has released continually updated [genomic surveillance data for Mycobacterium tuberculosis](#), the bacterium that causes tuberculosis (TB). TB is a major global health issue, causing more deaths around the world than any other infectious disease (WHO 2024).



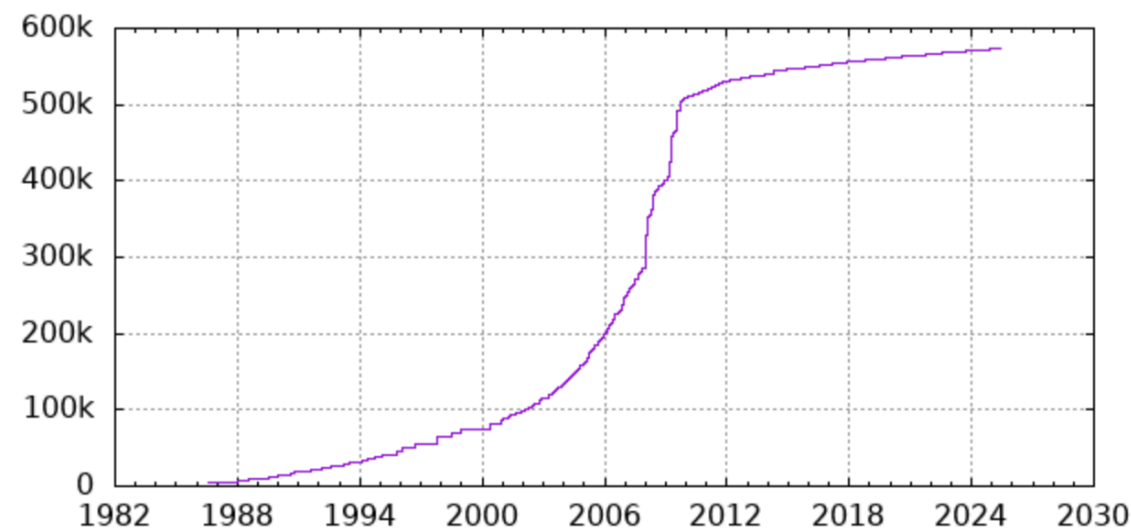
## Expasy

Swiss Bioinformatics Resource Portal



e.g. [BLAST](#), [UniProt](#), [MSH6](#), [Albumin...](#)

Number of entries in UniProtKB/Swiss-Prot



<http://www.expasy.org/>

# The function

Multifunctional transcription factor that induces cell cycle arrest, DNA repair or apoptosis upon binding to its target DNA sequence  
More than 10 publications

UniProt

BLASTAlignPeptide searchID mappingSPARQLUniProtKB

AdvancedListSearch

Help

Function

Names & Taxonomy

Subcellular Location

Disease & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

P04637

·

P53\_HUMAN

Protein<sup>i</sup>

Cellular tumor antigen p53

Amino acids

393 (go to sequence)

Gene<sup>i</sup>

TP53

Protein existence<sup>i</sup>

Evidence at protein level

Status<sup>i</sup>

UniProtKB reviewed (Swiss-Prot)

Annotation score<sup>i</sup>

5/5

Organism<sup>i</sup>

Homo sapiens (Human)

Entry

Variant viewer3,459

Feature viewer

Genomic coordinates

Publications

External links

History

Tools

Download

Add

Community curated (1)

Add a publication

Entry feedback

Function<sup>i</sup>

Multifunctional transcription factor that induces cell cycle arrest, DNA repair or apoptosis upon binding to its target DNA sequence (PubMed:11025664, PubMed:12524540, PubMed:12810724, PubMed:15186775, PubMed:15340061, PubMed:17317671, PubMed:17349958, PubMed:19556538, PubMed:20673990, PubMed:20959462, PubMed:22726440, PubMed:24051492, PubMed:24652652, PubMed:35618207, PubMed:36634798, PubMed:38653238, PubMed:9840937).  
Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type (PubMed:11025664, PubMed:12524540, PubMed:12810724, PubMed:15186775, PubMed:15340061, PubMed:17189187, PubMed:17317671, PubMed:17349958, PubMed:19556538, PubMed:20673990, PubMed:20959462, PubMed:22726440, PubMed:24051492, PubMed:24652652, PubMed:38653238, PubMed:9840937).  
Negatively regulates cell division by controlling expression of a set of genes required for this process (PubMed:11025664, PubMed:12524540, PubMed:12810724, PubMed:15186775, PubMed:15340061, PubMed:17317671, PubMed:17349958, PubMed:19556538, PubMed:20673990, PubMed:20959462, PubMed:22726440, PubMed:24051492, PubMed:24652652, PubMed:9840937).  
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 (PubMed:12524540, PubMed:17189187).

# 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.fasta)

[http://www.uniprot.org/uniprot/P53\\_HUMAN.txt](http://www.uniprot.org/uniprot/P53_HUMAN.txt)

More complex queries:

[http://www.uniprot.org/help/programmatic\\_access](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).
```



# Function & Computing

Can we transform functional annotation in computer readable information?

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

GENE ONTOLOGY  
Unifying Biology

About Ontology Annotations Downloads Help

PAN-GO Functionome: Working on human protein-coding genes? Click here to access the new PAN-GO Functionome!

Current release 2025-10-10: 39,354 GO terms | 9,281,704 annotations  
1,601,555 gene products | 5,495 species (see statistics)

## THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...

Any • Ontology • Gene Product

### GO Enrichment Analysis

Powered by PANTHER

Your gene IDs here...

biological process

Homo sapiens Examples Launch >

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs  
PAN-GO enrichment can be found here

## Ontology

Property	Value
Valid terms	39354 ( $\Delta = -552$ )
Obsoleted terms	8842 ( $\Delta = 586$ )
Merged terms	2436 ( $\Delta = 0$ )
Biological process terms	25153
Molecular function terms	10143
Cellular component terms	4058

## Annotations

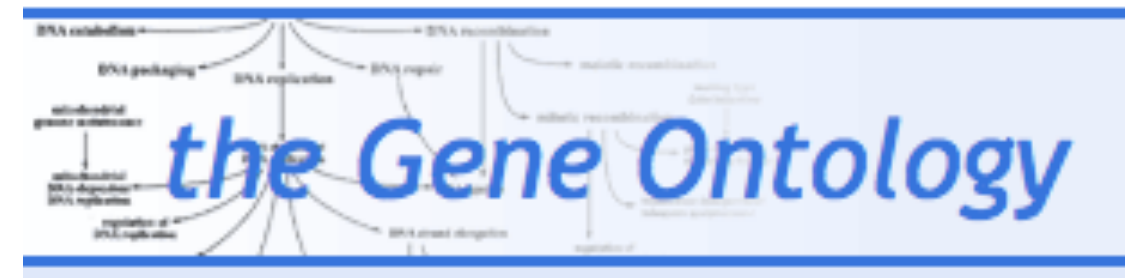
Property	Value
Number of annotations	9,281,704
Annotations for biological process	3,285,535
Annotations for molecular function	3,127,586
Annotations for cellular component	2,868,583
Annotations for evidence PHYLO	4,245,385
Annotations for evidence IEA	2,846,416
Annotations for evidence EXP	1,099,984
Annotations for evidence OTHER	940,851
Annotations for evidence ND	85,475
Annotations for evidence HTP	63,593
Number of annotated scientific publications	187,286

## Gene products and species

Property	Value
Annotated gene products	1,601,555
Annotated species	5,495
Annotated species with over 1,000 annotations	173

# 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

RCSB PDB Deposit Search Visualize Analyze Download Learn About Careers COVID-19 Help Contact us MyPDB

RCSB PDB PROTEIN DATA BANK 248,329 Structures from the PDB archive 1,068,577 Computed Structure Models (CSM)

Enter search term(s), Ligand ID or sequence Include CSM

Advanced Search | Chemical Search | Browse Annotations Help

PDB-101 PDB EMDDataResource NAKB wwPDB Foundation PDB-IHM

f X T Q in



Redesigned PDB Statistics Support Enhanced Functionality

Explore Statistics

Welcome

Deposit

Search

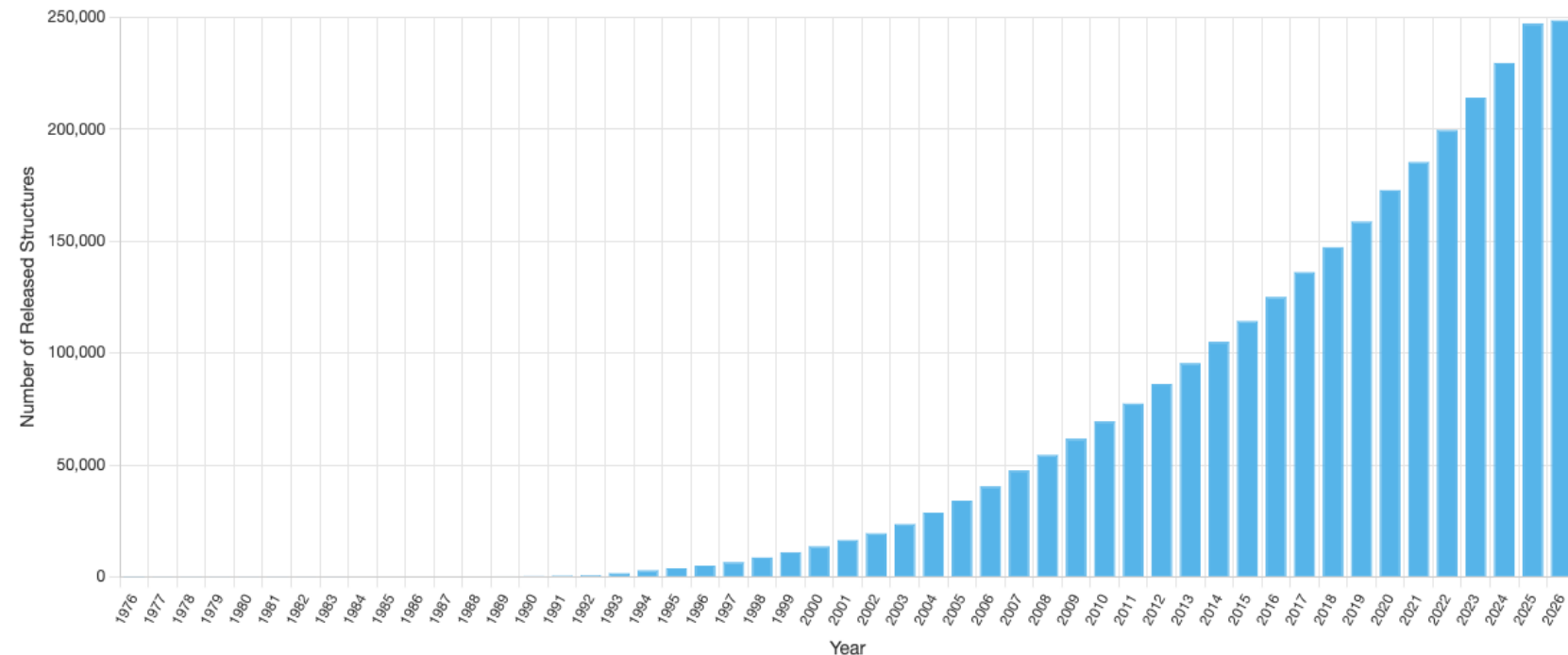
Visualize

Analyze

Download

Learn

PDB Data Growth by Released Structures



# 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  
[https://biofold.org/emidio/courses/vogelstein\\_tsg.txt](https://biofold.org/emidio/courses/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](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](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.