Biological Databases

Biology Is A Data Science

- Hundreds of thousand of species
- Million of articles in scientific literature
- Genetic Information
  - Gene names
  - Phenotype of mutants
  - Location of genes/mutations on chromosomes
  - Linkage (relationships between genes)

Data and Metadata

- Data are “concrete” objects
  - e.g. number, tweet, nucleotide sequence
- Metadata describes properties of data
  - e.g. object is a number, each tweet has an author
- Database structure *may* contain metadata
  - Type of object (integer, float, string, etc)
  - Size of object (strings at most 4 characters long)
  - Relationships between data (chromosomes *have* zero or more genes)

What is a Database?

- A data collection that needs to be :
  - Organized
  - Searchable
  - Up-to-date
- Challenge:
  - change “meaningless” data into useful, accessible information

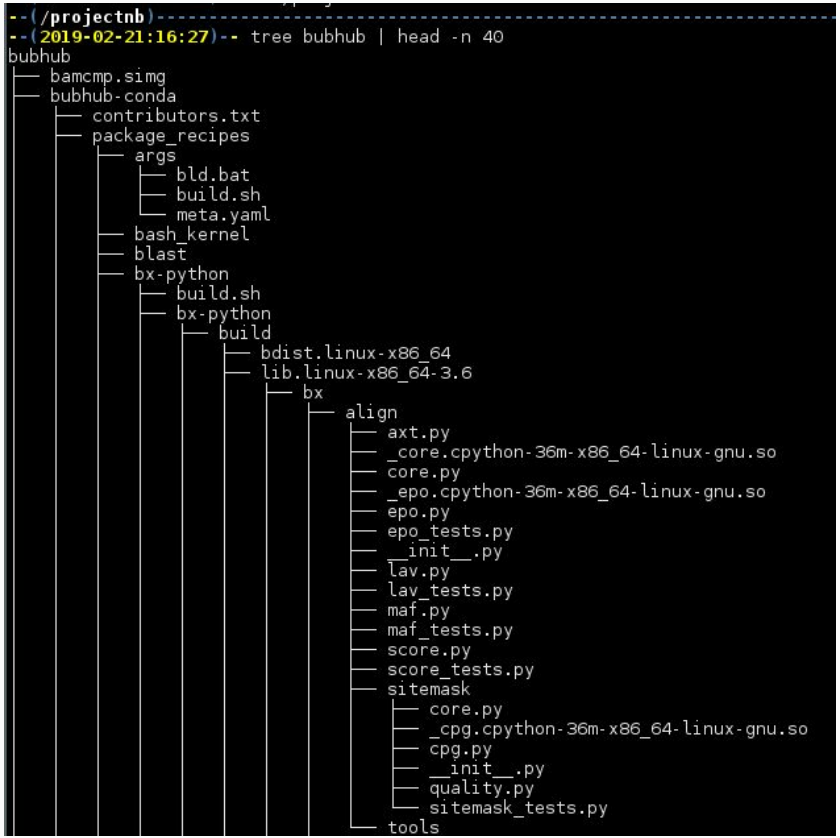
A spreadsheet can be a Database

- Rectangular data
- Structured
- No metadata
- Search tools:
  - Excel
  - `grep`
  - python/R

SNP ID	SNPSeq ID	Gene	+primer	-primer	Hap A	Hap B	Hap C
D1Mit160_1	10.MMHAP67FLD1.seq	lymphocyte antigen 84	AAGGTAAAAGGCAATCAGCACAGCC	TCAACCTGGAGTCAGAGGCT	C	—	A
M-05554_1	12.MMHAP31FLD3.seq	procollagen, type III, alpha	TGCGCAGAAGCTGAAGTCTA	TTTTGAGGTGTTAATGGTTCT	C	—	A
M-05554_2	X60184	complement component factor i	ACTTCCAGCCCTGGCTCT	ATATGCCACCAAGAAGCA	A	C	—
M-09947_3	AF067835	caspase 8	TCACAGAGGGAAACATGAAG	CTCCACATTGAACCAAAGCA	G	C	T
M-11415_1	U02023	insulin-like growth factor binding protein	GGGAAAAGCCTGAAAGAAGC	AGCTGAAACCGGACATCAAT	T	G	—
D1Mit284_3	J05234	nucleolin	TGTTGGAACCGACTTCTTCA	AAGAGTCAAAGAATTTATGGAATGA	G	T	T

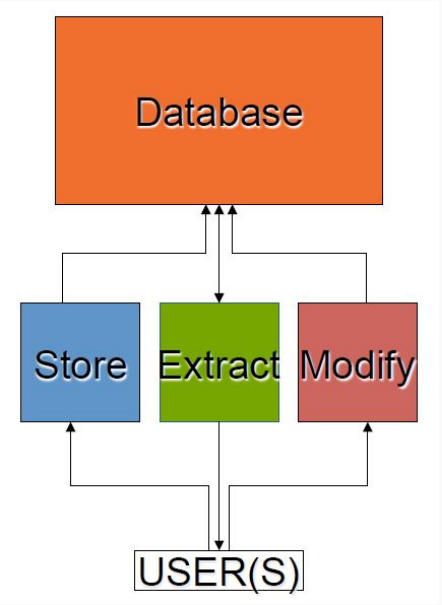
A filesystem can be a Database

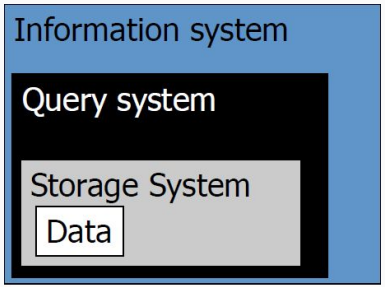
- Hierarchical data
- Some metadata
  - File, symlink, etc
- Unstructured
- Search tools:
  - `ls`
  - `find`
  - `locate`



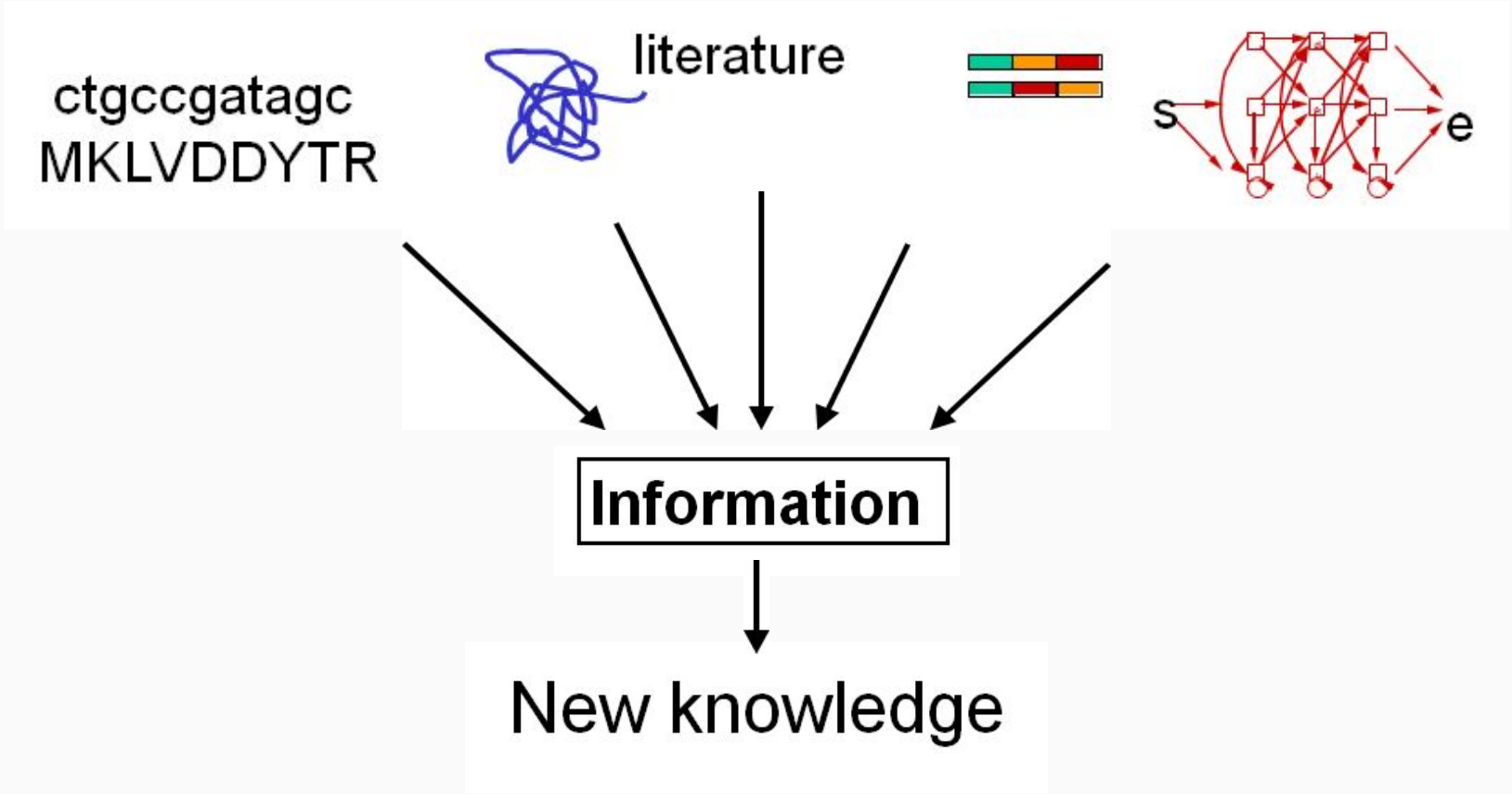
Organization and Types of Databases

- Every database has tools that:
  - Store
  - Extract
  - Modify
- Flat file databases (flat DBMS)
  - Simple, restrictive, table
- Hierarchical databases
  - Simple, restrictive, tables
- Relational databases (RDBMS)
  - Complex, versatile, tables
- Object-oriented databases (ODBMS)
- Data warehouses and distributed databases
- Unstructured databases (object store DBs)





Where do the data come from ?

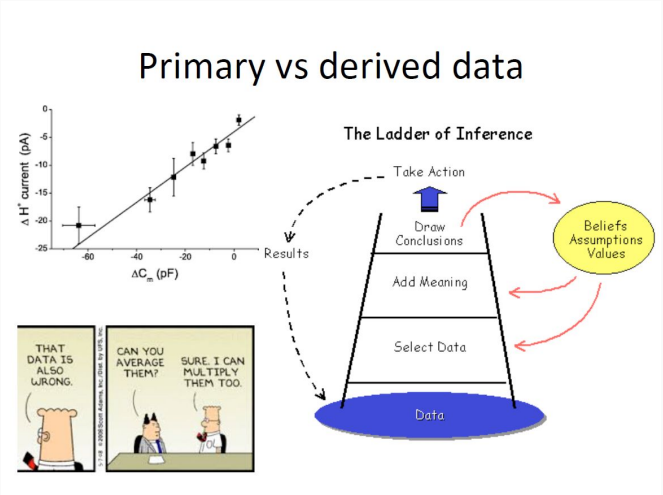


Types of Biological Data

- Primary data types (observed properties):
  - Molecular Sequence: nucleic or amino acids
  - Quantity: DNA, RNA, Protein, cell count, metabolites
  - Locality: membrane, nucleus, epithelium
  - Structure: 3D conformation, proximity, size
- Secondary data types (inferred properties):
  - Molecular Function
  - Dynamics
  - Relation: multiplicity, distribution, binding
  - Association: co-occurrence, correlation
  - Predicted: computational models

Types of Biological Databases

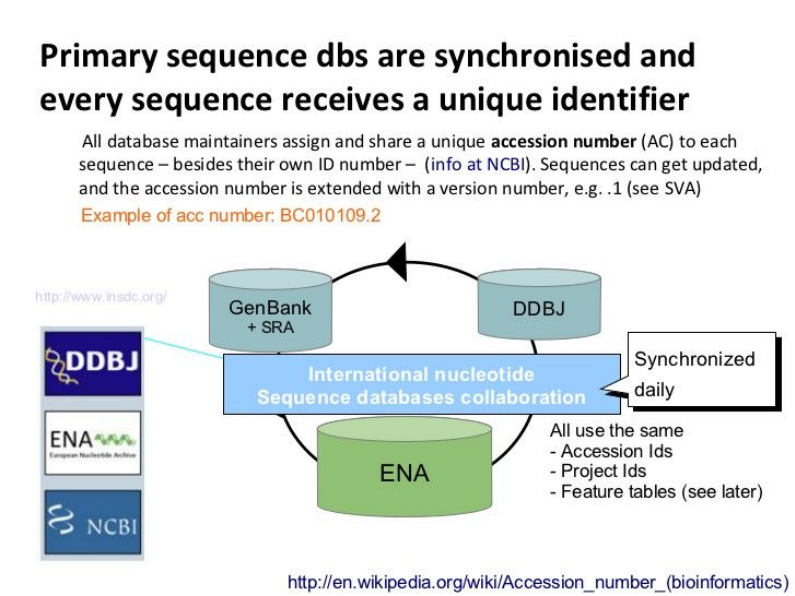
- **Primary Databases:**
  - Original submissions by experimentalists
  - Content controlled by the **submitter**
  - Examples: GenBank, Trace, SRA, SNP, GEO
- **Secondary databases:**
  - Results of analysis of primary databases
  - Aggregate of many databases
  - Content controlled by **third party** (NCBI)
  - Examples: NCBI Protein, Refseq, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain



NCBI

- National Center for Biotechnology Information
- >50 different biological databases and tools
- Most popular:
  - PubMed - search for citations/papers
  - BLAST - search for sequences
  - Nucleotide - all nucleotide sequences (DNA, etc)
  - Genome - published genomes
  - SNP (dbSNP) - all human SNPs
  - Protein - amino acid sequences

International Sequence Database Collaboration



International Sequence Database Collaboration: http://www.insdc.org/
National Centre for Biotechnology Information (NCBI) : https://www.ncbi.nlm.nih.gov/
European Nucleotide Archive (ENA) : https://www.ebi.ac.uk/ena
DNA Data Bank of Japan (DDBJ) : http://www.ddbj.nig.ac.jp/

Data sharing collaboration

Data type	Collaboration
Nucleotide sequences	International Sequence Database Collaboration
Protein sequences	UniProt Consortium
Macromolecular structures	Worldwide Protein Data Bank
Molecular interactions	The International Molecular Exchange Consortium
Protein identifications	The ProteomeXchange Consortium
Metabolomics data	Coordination of Standards in Metabolomics
Genomic and clinical data	Global Alliance for Genomics and Health

- Ensure data consistency
- Avoid duplication
- Open data sharing

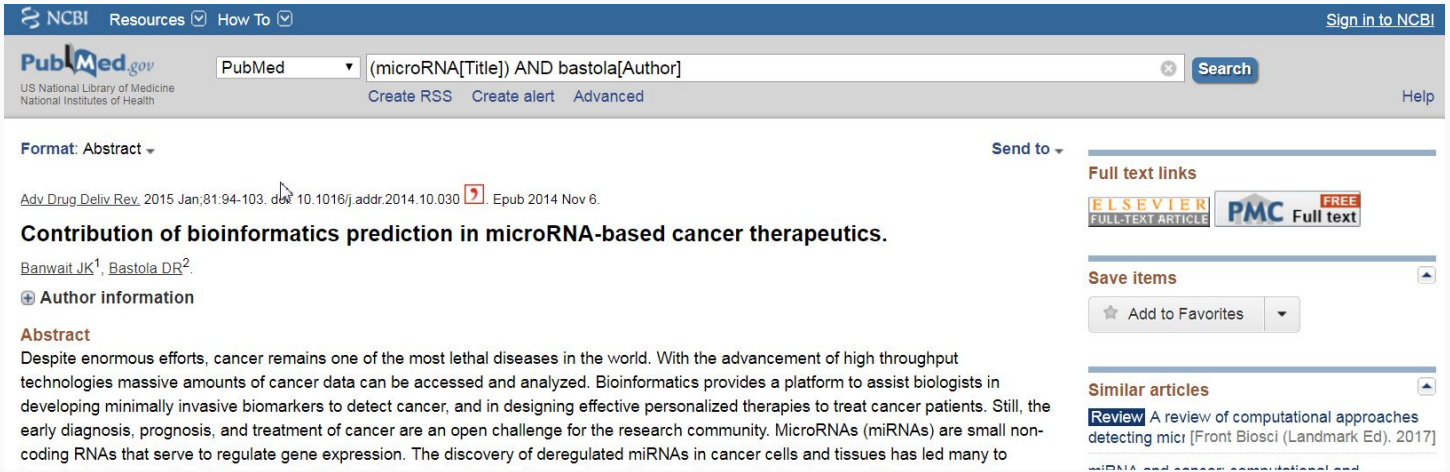
Biological Databases I: Biomedical Literature

Biological Database I : Biomedical Literature Database

- Citation/publication databases
- Medline:
  https://www.nlm.nih.gov/bsd/pmresources.html
  - NLM journal citation database.
  - Includes citations 5,600 scholarly journals
- PubMed
  https://www.ncbi.nlm.nih.gov/pubmed/
    - Includes MEDLINE
    - journals/manuscripts deposited in PMC
    - NCBI Bookshelf

Searching PubMed with MeSH terms

- **MeSH (Medical Subject Headings)** is the NLM controlled vocabulary used for indexing articles for PubMed.
  - the U.S. National Library of Medicine's controlled vocabulary
  - arranged in a hierarchical manner called the MeSH Tree Structures
  - updated annually



Biological Databases II: Genomics and Transcriptomics

Biological Database II - Genomics and Transcriptomics

- **GenBank: https://www.ncbi.nlm.nih.gov/genbank/**
  - **Flat file**
  - **DNA only** sequence database
  - **Archival** in nature: Historical, Redundant
  - Sample GenBank record (accession number **U49845**)
    - NCBI: https://www.ncbi.nlm.nih.gov/nuccore/U49845
    - ENA: https://www.ebi.ac.uk/ena/data/view/U49845
    - DDBJ: http://getentry.ddbj.nig.ac.jp/top-e.html

GenBank Flat File

```
LOCUS       MUSNGH      1803 bp    mRNA            ROD       29-AUG-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
NID         g1850791
KEYWORDS    neurite extension activity; growth arrest; TA20.
SOURCE      Murinae gen. sp. mouse neuroblastma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
  ORGANISM  Murinae gen. sp.
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE   1  (sites)
  AUTHORS   Tohda,C., Nagai,S., Tohda,M. and Nomura,Y.
  TITLE     A novel factor, TA20, involved in neuronal differentiation: cDNA
            cloning and expression
  JOURNAL   Neurosci. Res. 23 (1), 21-27 (1995)
  MEDLINE   96064354
REFERENCE   3  (bases 1 to 1803)
  AUTHORS   Tohda,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-NOV-1993) to the DDBJ/EMBL/GenBank databases. Chihiro
            Tohda, Toyama Medical and Pharmaceutical University, Research
            Institute for Wakan-yaku, Analytical Research Center for
            Ethnomedicines; 2630 Sugitani, Toyama, Toyama 930-01, Japan
            (E-mail:CHIHIRO@ms.toyama-mpu.ac.jp, Tel:+81-764-34-2281(ex.2841),
            Fax:+81-764-34-5057)
COMMENT     On Feb 26, 1997 this sequence version replaced gi:793764.
FEATURES             Location/Qualifiers
     source          1..1803
                     /organism="Murinae gen. sp."
                     /note="source origin of sequence, either mouse or rat, has
                     not been identified"
                     /db_xref="taxon:39108"
                     /cell_line="NG108-15"
                     /cell_type="mouse neuroblastma-rat glioma hybridoma"
     misc_signal     156..163
                     /note="AP-2 binding site"
     GC_signal       647..655
                     /note="Sp1 binding site"
     TATA_signal     694..701
     gene            748..1311
                     /gene="TA20"
     CDS             748..1311
                     /gene="TA20"
                     /function="neurite extension activity and growth arrest
                     effect"
                     /codon_start=1
                     /db_xref="PID:d1005516"
                     /db_xref="PID:g793765"
                     /translation="MMKLWVPSRSLFNSPNHYRSFLSHTLHIRYNNSLFISNTHLSRR
                     KLRVTNPIYTRKRSLNIFYLLIPSCRTRLILWIIVIYRNLKHWSTSTVRSHSRSIYRL
                     RPSMRTWILLRCHSVYKPPISNFIYWNNPSRWMNLRGLLSRQSHLDPILRFPLHLTIYY
                     RGPSNRSPPLPPNNRIKQPNRIKLRCR"
     polyA_site      1803
BASE COUNT      507 a    458 c    311 g    527 t
ORIGIN
        1 tcagtttttt tttttttttt tttttttttt tttttttttt tttttttttg ttgattcatg
       61 tccgtttaca tttggtaagt tcacaggcct cagtcaacac aattggactg ctcaggaaat
      121 cctccttggt gaccgcagta tacttggcct atgaacccaa gccacctatg gctaggtagg
      181 agaagctcaa ctgtagggct gactttggaa gagaatgcac atggctgtat cgacatttca
      241 catggtggac ctctggccag agtcagcagg ccgaggggttc tcttccgggc tgctccctca
      301 ctgcttgact ctgcgtcagt gcgtccatac tgtgggcgga cgttattgct atttgccttc
      361 cattctgtac ggcattgcct ccatttagct ggagagggac agagcctggt tctctagggc
      421 gtttccattg gggcctggtg acaatccaaa agatgagggc tccaaacacc agaatcagaa
      481 ggcccagcgt atttgtaaaa acacctttg gtgggaatga atggtacagg ggcgttttag
      541 gacaaagaac agcttttctg tcactcccat gagaacgtgc gcaatcactg ttccgaagag
      601 gaggagtcca gaatacacgt gtatgggcat gacgattgcc cggagagagg cggagcccat
      661 ggaagcagaa agacgaaaaa cacacccatt atttaaaatt attaaccact cattcattga
      721 cctacctgcc ccatccaaca tttcatcatg atgaaacttt gggtcccttc taggagtctg
      781 cctaatagtc caaatcatta caggtctttt cttagccata cactacacat cagatacaat
      841 aacagccttt 1801 cat
//
```

Header

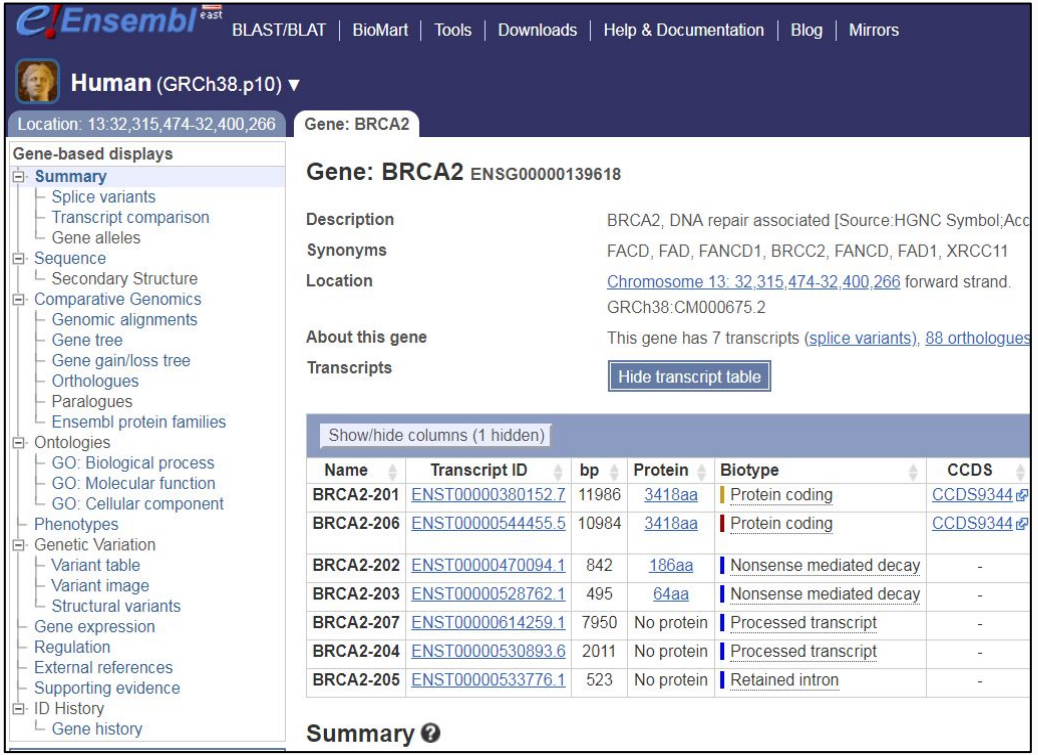
- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

Ensembl

- Comprehensive DNA/RNA sequence and annotation database
- Automated annotation:
  - Genes (known or predicted)
  - Single nucleotide polymorphisms (SNPs)
  - Repeats
  - Homology
- Analysis tools:
  - BLAST
  - BioMart
  - Variant Effect Predictor



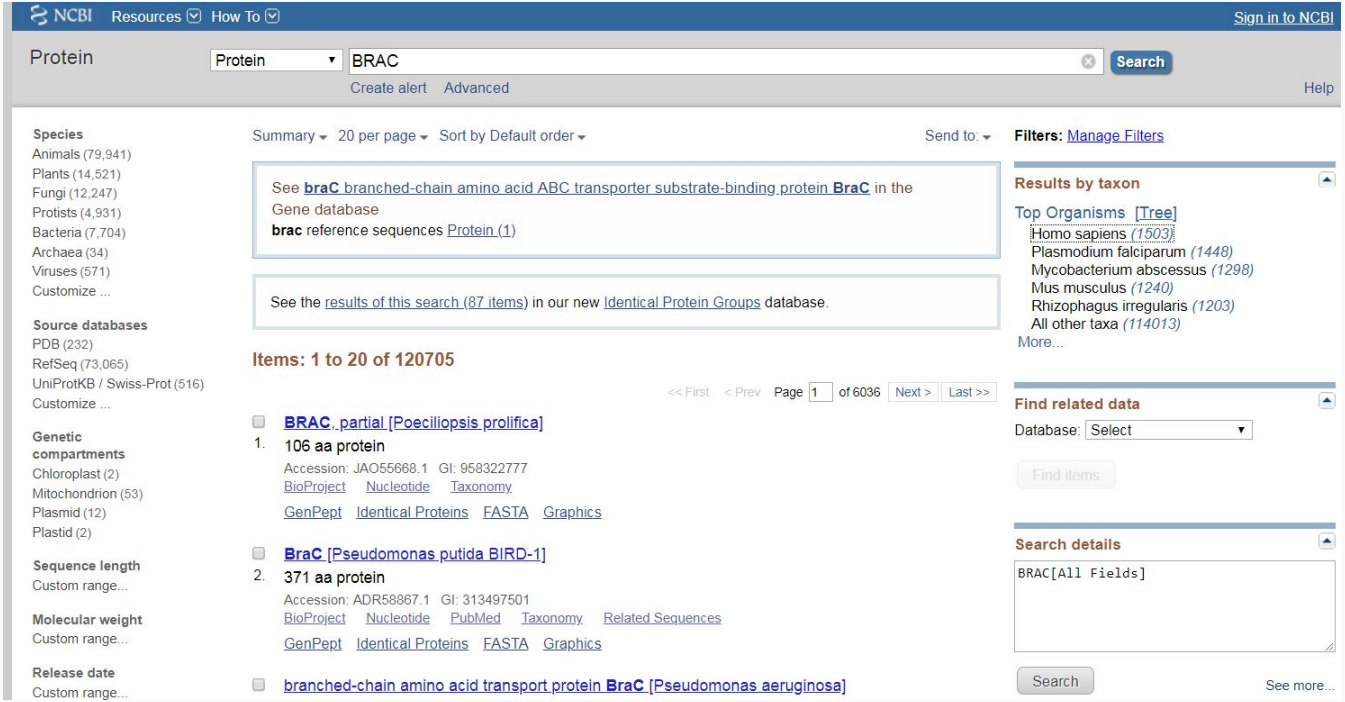
Nucleic Acid Structure Database

- NDB Nucleic acid-containing structures http://ndbserver.rutgers.edu/
- NTDB Thermodynamic data for nucleic acids http://ntdb.chem.cuhk.edu.hk/
- RNABase RNA-containing structures from PDB and NDB http://www.rnabase.org/
- SCOR Structural classification of RNA: RNA motifs by structure, function and tertiary interactions http://scor.lbl.gov/

Biological Databases III: Proteomics

Biological Database III - Proteomics

- Protein sequence database: https://www.ncbi.nlm.nih.gov/protein/



Genpept

Chain A, Structure Of The Btb (tramtrack And Bric A Gigaxonin

PDB: 2PPI_A

Identical Proteins FASTA Graphics

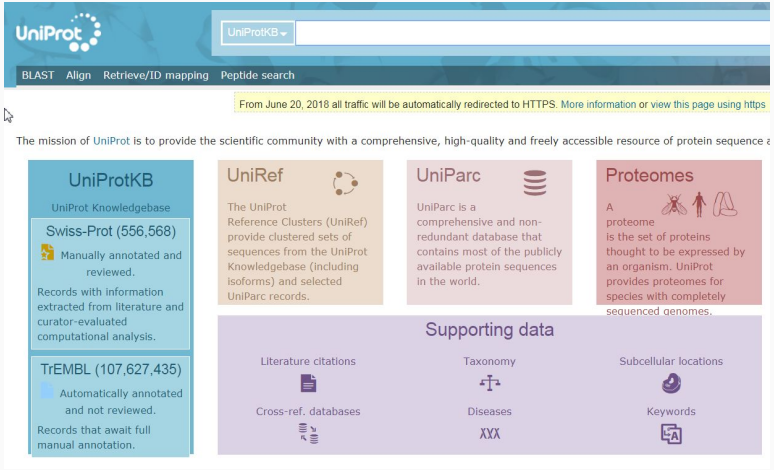
Go to:

```
LOCUS       2PPI_A                  144 aa            linear   PRI 26-OCT-20
DEFINITION  Chain A, Structure Of The Btb (tramtrack And Bric A Brac) Domain
            Human Gigaxonin.
ACCESSION   2PPI_A
VERSION     2PPI_A
DBSOURCE    pdb: molecule 2PPI, chain 65, release Oct 22, 2017;
            deposition: Apr 30, 2007;
            class: Structural Protein;
            source: Mmdb_id: 46639, Pdb_id 1: 2PPI;
            Exp. method: X-Ray Diffraction.
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 144)
  AUTHORS   Amos,A., Turnbull,A.P., Tickle,J., Keates,T., Bullock,A., S,P.,
            Burgess-Brown,N., Debreczeni,J.E., Ugochukwu,E., Umeano,C.,
            Pike,A.C.W., Papagrigoriou,E., Sundstrom,M., Arrowsmi,C.H.,
            Weigelt,J., Edwards,A., Von Delft,F. and Knapp,S.
  TITLE     Structure Of The Btb (Tramtrack And Bric A Brac) Do Human Gigaxor
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 144)
  AUTHORS   Amos,A., Turnbull,A.P., Tickle,J., Keates,T., Bullock,A.,
            Savitsky,P., N, Brown, Debreczeni,J.E., Ugochukwu,E., Umeano,C.,
            Pike,A.C.W., Papagrigoriou,E., Sundstrom,M., Arrowsmith,C.H.,
            Weigelt,J., Edwa,A., Delft, Knapp,S. and Structural Genomics
            Consortium (Sgc).
```

```
                     /note="BTB/POZ domain; pfam00651"
                     /db_xref="CDD:279045"
     Region          46..96
                     /region_name="Domain 2"
                     /note="NCBI Domains"
     Region          49..144
                     /region_name="BTB"
                     /note="Broad-Complex, Tramtrack and Bric a brac;
                     smart00225"
                     /db_xref="CDD:197585"
     SecStr          49..55
                     /sec_str_type="sheet"
                     /note="strand 1"
     SecStr          56..62
                     /sec_str_type="sheet"
                     /note="strand 2"
     SecStr          63..70
                     /sec_str_type="helix"
                     /note="helix 2"
     SecStr          71..79
                     /sec_str_type="helix"
                     /note="helix 3"
     SecStr          89..94
                     /sec_str_type="sheet"
                     /note="strand 3"
     SecStr          98..109
                     /sec_str_type="helix"
                     /note="helix 4"
     SecStr          120..130
                     /sec_str_type="helix"
                     /note="helix 5"
     SecStr          133..141
                     /sec_str_type="helix"
                     /note="helix 6"
ORIGIN
        1 mhhhhhhssg vdlgtenlyf qsmavsdpqh aarllralss freesrfcda hlvldgeeip
       61 vqknilaaas pyirtklnyn ppkddgstyk ielegisvmv mreildyifs gqirlnedti
      121 qdvvqaadll lltdlktlcc efle
//
```

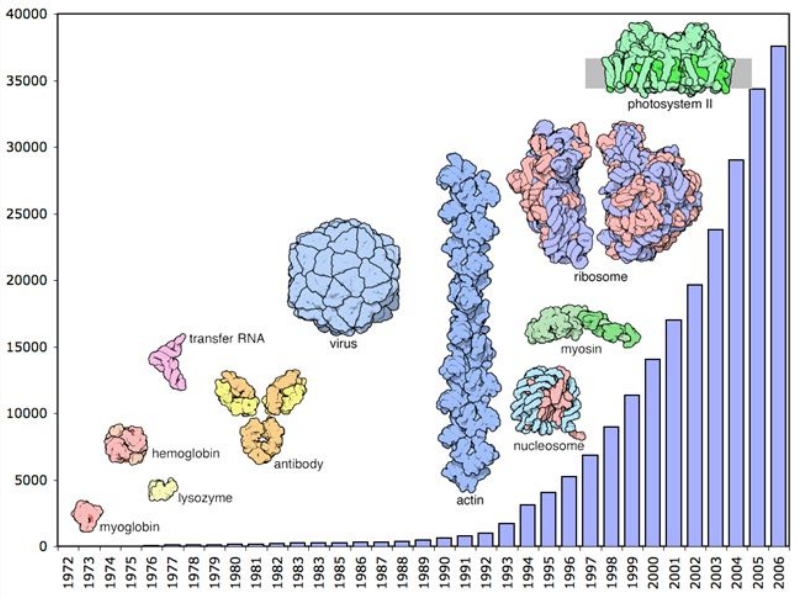
Uniprot

- The Universal Protein Resource
- Comprehensive resource for protein sequence and annotation data
- Collaboration between:
  - EMBL-EBI
  - Swiss Institute of Bioinformatics
  - Protein Information Resource
- Entries in two categories:
  - Swiss-Prot (experimentally verified)
  - TrEMBL (computer-annotated)
- http://www.uniprot.org/



Protein Structure database - PDB

- Protein Data Bank (PDB) http://www.rcsb.org/
- Dedicated to 3D structure of proteins and peptides
- ~150,000 predicted and experimental (solved) structures



PDB: kinesin 6

RCSB PDB PROTEIN DATA BANK

149174 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

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

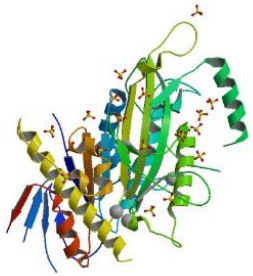
Go

Advanced Search | Browse by Annotations

PDB-101 | Worldwide PDB | EMDataResource | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers
Protein Workshop | Ligand Explorer

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A

Biological assembly 1 assigned by authors and generated by PISA (software)

Display Files | Download Files

5X3E

kinesin 6

DOI: 10.2210/pdb5X3E/pdb

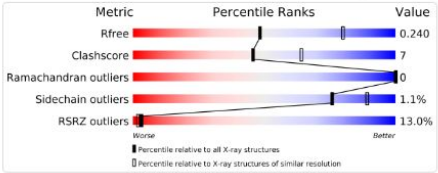
Classification: MOTOR PROTEIN
Organism(s): Caenorhabditis elegans
Expression System: Escherichia coli-Thermus thermophilus shuttle vector pTRH1T

Deposited: 2017-02-04 Released: 2017-04-19
Deposition Author(s): Chen, Z., Guan, R., Zhang, L.
Funding Organization(s): Chinese Key Research Plan-Protein Sciences; National Natural Science Foundation of China; Junior One Thousand Talents program; National Science Foundation of China; 863 Program

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.61 Å
R-Value Free: 0.240
R-Value Work: 0.209

wwPDB Validation | 3D Report | Full Report



This is version 1.0 of the entry. See complete history.

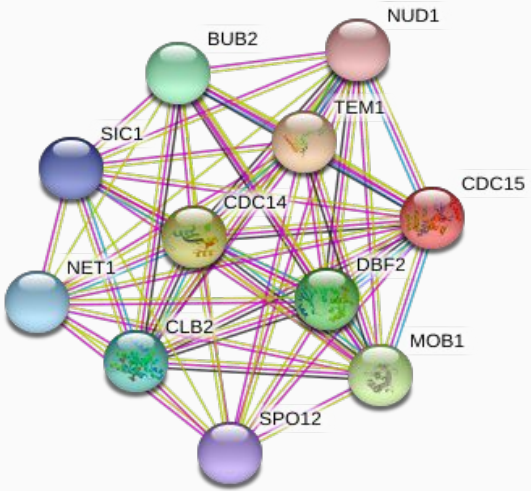
Literature | Download Primary Citation

Protein Family Database

- http://pfam.xfam.org/family/piwi
- Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models

Protein-Protein Interaction Database

- **STRING:** https://string-db.org/
- *Search Tool for the Retrieval of Interacting Genes/Proteins*
- Database of protein/protein interactions
- Information from numerous sources:
  - experimental data
  - computational prediction methods
  - public text collections
- Expressed as interaction graphs:
  - **Nodes:** Network nodes represent proteins
  - **Edges:** Edges represent protein-protein associations



Data vs Annotation Database

- **RefSeq**: curated nonredundant biological sequences https://www.ncbi.nlm.nih.gov/refseq/
  - Source: Genbank (INSDC)
  - Annotated: Community collaboration, automated computer, NCBI staff curation
- Advantages of using RefSeq
  - **Non-redundancy**
  - **Curated, validated**
  - **Format consistency**
  - **Distinct accession series**
  - **Updates to reflect current sequence data and biology**

Selected Refseq Accession

mRNAs and Proteins	
NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
Gene Records	
NG_123456	Reference Genomic Sequence
Chromosome	
NC_123455	Microbial replicons, organelle
AC_123455	Alternate assemblies
Assemblies	
NT_123456	Contig
NW_123456	WGS Supercontig

High-Throughput Sequencing Databases

- Gene Expression Omnibus (GEO)
- NCBI Sequence Read Archive (SRA)
- db of Genotype and Phenotype (dbGAP)
- European Genome Phenome Archive

Other Specialised Databases

- UCSC Xena: https://xenabrowser.net/datapages/
- Genotype-Tissue Expression Gtex: https://www.gtexportal.org/home/
- mirBase:http://www.mirbase.org/
- Pubchem: https://pubchem.ncbi.nlm.nih.gov/
- DrugBank: https://www.drugbank.ca/
- Many more...

BLAST

- BLAST stands for Basic Local Alignment Search Tool
  - Good balance of sensitivity and speed
  - Searches all publicly available sequences
  - Flexible
- Produce local alignments
- Applies heuristic, no optimality guarantee

Program	Database (Subject)	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. → Protein
TBLASTN	Nt. → Protein	Protein
TBLASTX	Nt. → Protein	Nt. → Protein

NAR Database Issue

- Online collection of biological databases:

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