

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	AAGGTAAA GGCAATCAG CACAGCC	TCAACCTGG AGTCAGAGG CT	C	—	A
M-05554_1	12.MMHAP31FLD3.seq	procollagen, type III, alpha	TGCGCAGAA GCTGAAGTC TA	TTTTGAGGT GTTAATGGTT CT	C	—	A
M-05554_2	X60184	complement component factor i	ACTTCCAGC CCTGGCTCT	ATATGCCACC AAGAAGCA	A	C	—
M-09947_3	AF067835	caspase 8	TCACAGAGG GAAACATGA AG	CTCCACATTG AACCAAAGC A	G	C	T
M-11415_1	U02023	insulin-like growth factor binding protein	GGGAAAAGC CTGAAAGAA GC	AGCTGAAAC CGGACATCA AT	T	G	—
D1Mit284_3	J05234	nucleolin	TGTTGGAAC CGACTTCTTC A	AAGAGTCAA AGAATTTATG GAATGA	G	T	T

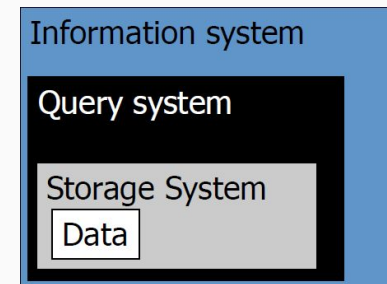
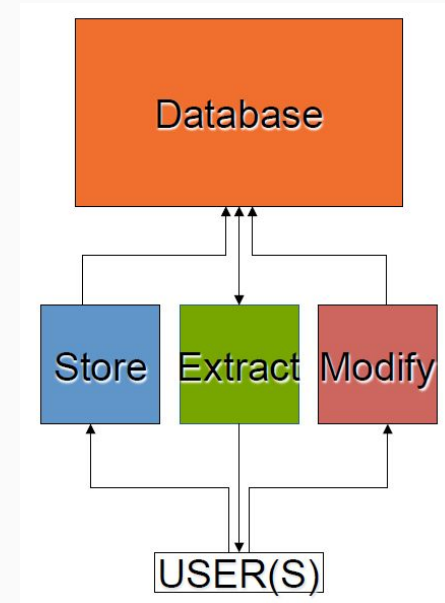
A filesystem can be a Database

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

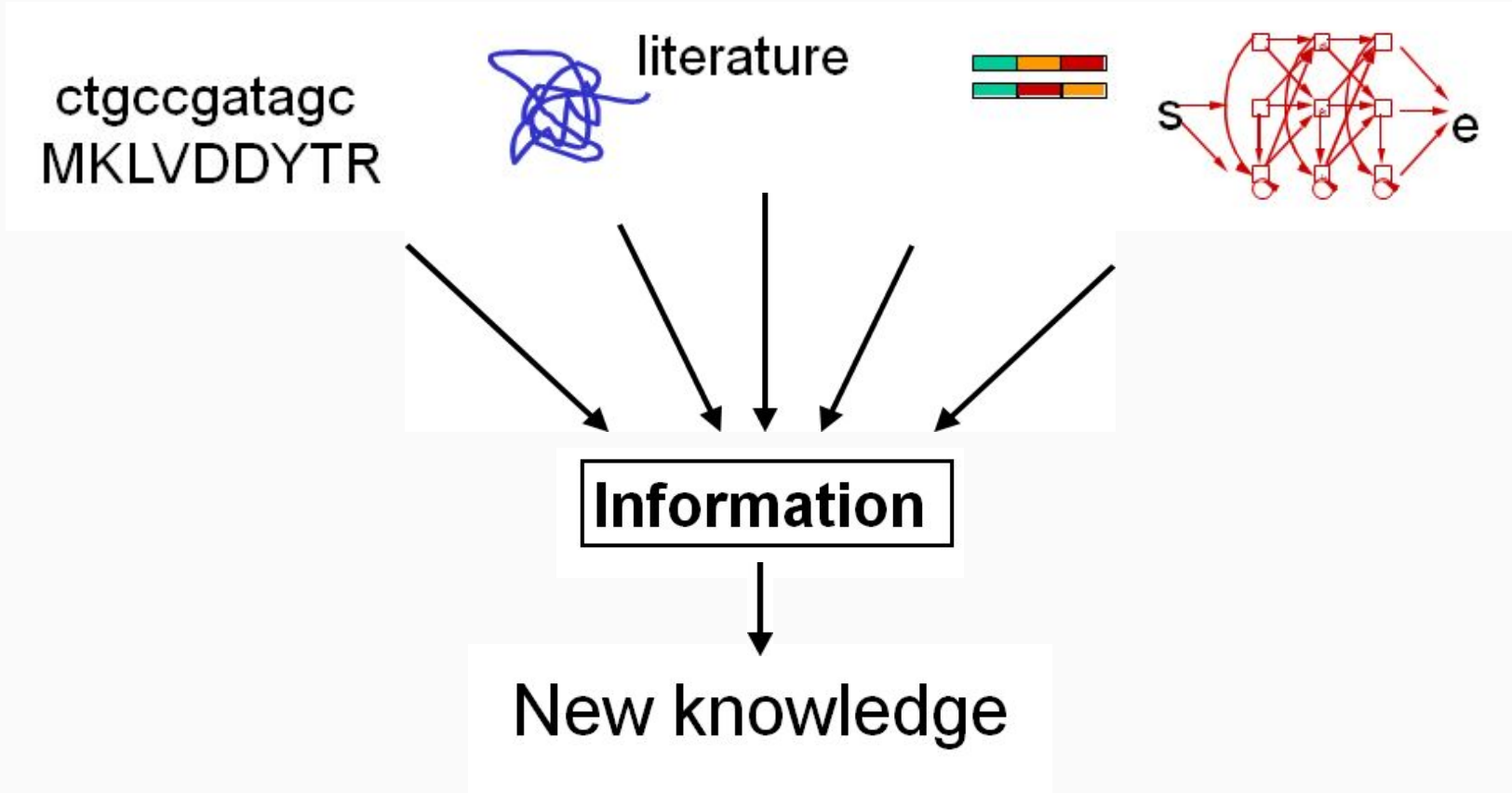
```
--(/projectnb)
--(2019-02-21:16:27)-- tree bubhub | head -n 40
bubhub
├── bamcmp.simg
├── bubhub-conda
│   ├── contributors.txt
│   ├── package_recipes
│   │   ├── args
│   │   │   ├── bld.bat
│   │   │   ├── build.sh
│   │   │   └── meta.yaml
│   │   ├── bash_kernel
│   │   ├── blast
│   │   └── bx-python
│   │       ├── build.sh
│   │       └── bx-python
│   │           ├── build
│   │           ├── bdist.linux-x86_64
│   │           └── lib.linux-x86_64-3.6
│   │               └── bx
│   │                   └── align
│   │                       ├── axt.py
│   │                       ├── _core.cpython-36m-x86_64-linux-gnu.so
│   │                       ├── core.py
│   │                       ├── _epo.cpython-36m-x86_64-linux-gnu.so
│   │                       ├── epo.py
│   │                       ├── epo_tests.py
│   │                       ├── _init_.py
│   │                       ├── lav.py
│   │                       ├── lav_tests.py
│   │                       ├── maf.py
│   │                       ├── maf_tests.py
│   │                       ├── score.py
│   │                       ├── score_tests.py
│   │                       ├── sitemask
│   │                       ├── core.py
│   │                       ├── _cpg.cpython-36m-x86_64-linux-gnu.so
│   │                       ├── cpg.py
│   │                       ├── _init_.py
│   │                       ├── quality.py
│   │                       ├── sitemask_tests.py
│   │                       └── tools
```

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

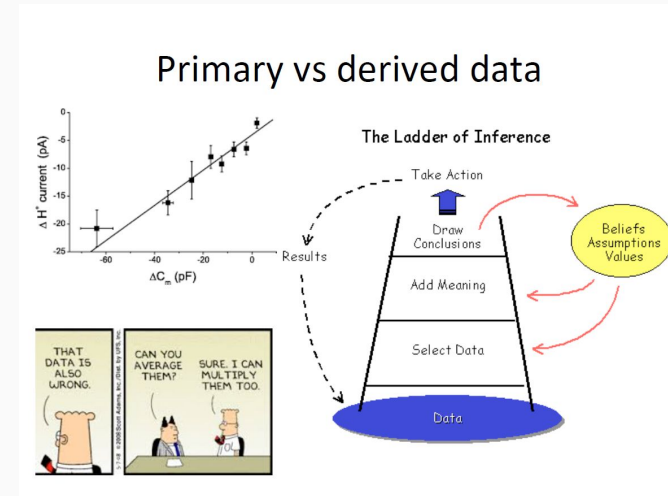
Types of Biological Databases

- **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

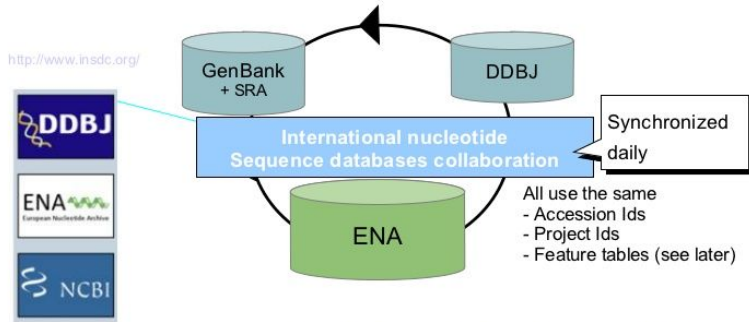
- National Center for Biotechnology Information
- >50 different biological databases and tools
- Most popular:
 - PubMed - search for citations/papers
 - BLAST - search for sequences
 - Nucleotide - all nucleotide sequences (DNA, etc)
 - Genome - published genomes
 - SNP (dbSNP) - all human SNPs
 - Protein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

International Sequence Database Collaboration

Primary sequence dbs are synchronised and every sequence receives a unique identifier

All database maintainers assign and share a unique **accession number** (AC) to each sequence – besides their own ID number – (info at NCBI). Sequences can get updated, and the accession number is extended with a version number, e.g. .1 (see SVA)

Example of acc number: **BC010109.2**



[http://en.wikipedia.org/wiki/Accession_number_\(bioinformatics\)](http://en.wikipedia.org/wiki/Accession_number_(bioinformatics))

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

- 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

The screenshot shows the PubMed search interface. At the top, there is a search bar with the query "(microRNA[Title]) AND bastola[Author]" and a "Search" button. Below the search bar, the article title "Contribution of bioinformatics prediction in microRNA-based cancer therapeutics." is displayed, along with the authors "Banwait JK¹, Bastola DR²". The abstract text is visible, starting with "Despite enormous efforts, cancer remains one of the most lethal diseases in the world. With the advancement of high throughput technologies massive amounts of cancer data can be accessed and analyzed. Bioinformatics provides a platform to assist biologists in developing minimally invasive biomarkers to detect cancer, and in designing effective personalized therapies to treat cancer patients. Still, the early diagnosis, prognosis, and treatment of cancer are an open challenge for the research community. MicroRNAs (miRNAs) are small non-coding RNAs that serve to regulate gene expression. The discovery of deregulated miRNAs in cancer cells and tissues has led many to". On the right side of the page, there are sections for "Full text links" (including Elsevier and PMC Full text), "Save items" (with an "Add to Favorites" button), and "Similar articles" (with a "Review" link).

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       MUSNRH             1803 bp    mRNA             ROD             29-AGO-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
RID        gl850791
KEYWORDS   neurite extension activity; growth arrest; TA20.
SOURCE     Murinae gen. sp. mouse neuroblastoma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
ORGANISM   Murinae gen. sp.
            Eukaryotes; mitochondrion eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE  1 (bases)
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    9604354
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; 2430 Sogitani, Toyama, Toyama 930-01, Japan
            (E-mail:CHRITIRO@ms.toyama-npu.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 neuroblastoma-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="MAGLWVPSRSLPNSNPNVRSFLRSLIRVWVNSLFIENSLRSLR
            KLRVNFIVYKRSINIFYLLPQRTSLILMIVYVRLKSNSTVSRSHSISYLR
            RPEMRTNIIKCSYVYFPIIHPVWVNSRNRLGLLRSRQSHLOPLRFLNLTIVY
            RPSNRSFPLFPNRIRIQNRIKLRK"
            polyA_site
            1803
BASE COUNT  507 a   458 c   311 g   527 t
ORIGIN
1  ccagttttt  tttttttt  tttttttt  tttttttt  tttttttt  ttgattcatg
61  tccgtttaca  ttgttaagt  tcaacggct  cagtaaacac  aattggactg  ctacaggaat
121  cctccttggt  gaccgcagta  tacttggct  atgaccocaa  gccacctatg  gctaggtagg
181  agaaactcaa  ctgtagggt  gactttgga  gagaatgac  atgctgtat  cagactttca
241  catgtggac  ctctccag  agtcagcag  ccagaggctc  tcttcaggcg  tgcctcaata
301  ctgattgact  ctgctcag  gptccatcc  tgtggggaga  ogttattgct  attgcttc
361  cattctgtac  ggcattgct  ccatttagc  ggagagggac  agagcttgg  tototagggc
421  gtttccattg  gggctggg  acatoccaa  agatgaggg  tcaaacacac  agaatcaga
481  ggcocaggt  attgtaaa  acactctgt  gtggatga  atgtacag  gggcttcag
541  gacaagaac  agctttctg  tcactccat  gagaacctc  gaactcag  ttcogaag
601  gaggatcca  gaatacagt  gtatggcat  gacatggcc  cggagaggg  cggagcccat
661  ggaagcaga  agagaana  cacaccat  atttaaat  attaacat  catctatg
721  ctcaactg  caatcaaa  tttcactag  atgaacct  gggcttc  taggctg
781  cctaactg  caatcatta  naggtttt  cttagccata  caactacat  nagatcat
841  aacagcttt  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

The screenshot displays the Ensembl genome browser interface for the BRCA2 gene. The top navigation bar includes links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. The main header shows the species as Human (GRCh38.p10) and the gene as BRCA2. The left sidebar contains a 'Gene-based displays' menu with categories like Summary, Sequence, Comparative Genomics, Ontologies, and Genetic Variation. The main content area shows the gene's description, synonyms, location, and a table of transcripts. The transcript table has columns for Name, Transcript ID, bp, Protein, Biotype, and CCDS. Below the table is a 'Summary' section.

Human (GRCh38.p10)
Location: 13:32,315,474-32,400,266
Gene: BRCA2

Gene: BRCA2 ENSG00000139618

Description BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:...]
Synonyms FACD, FAD, FANCD1, BRCC2, FANCD, FAD1, XRCC11
Location [Chromosome 13: 32,315,474-32,400,266](#) forward strand.
GRCh38:CM000675.2

About this gene This gene has 7 transcripts ([splice variants](#)), [88 orthologues](#)

Transcripts [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS
BRCA2-201	ENST00000380152.7	11986	3418aa	Protein coding	CCDS9344.4
BRCA2-206	ENST00000544455.5	10984	3418aa	Protein coding	CCDS9344.4
BRCA2-202	ENST00000470094.1	842	186aa	Nonsense mediated decay	-
BRCA2-203	ENST00000528762.1	495	64aa	Nonsense mediated decay	-
BRCA2-207	ENST00000614259.1	7950	No protein	Processed transcript	-
BRCA2-204	ENST00000530893.6	2011	No protein	Processed transcript	-
BRCA2-205	ENST00000533776.1	523	No protein	Retained intron	-

Summary ⓘ

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/>

The screenshot shows the NCBI Protein database search results for the query 'BRAC'. The interface includes a search bar at the top with 'Protein' selected as the database and 'BRAC' as the search term. Below the search bar, there are navigation options like 'Create alert' and 'Advanced', and a 'Help' link. The main content area displays search results for 'BRAC', including a summary of the search (87 items) and a list of items. The first item is 'BRAC [Pseudomonas putida BIRD-1]', a 371 aa protein with accession number ADR58867.1. The second item is 'BRAC [Pseudomonas aeruginosa]', a branched-chain amino acid transport protein. The interface also features a sidebar with filters for Species, Source databases, Genetic compartments, Sequence length, Molecular weight, and Release date. On the right side, there are sections for 'Results by taxon' (listing top organisms like Homo sapiens and Plasmodium falciparum) and 'Find related data' (with a dropdown menu for Database).

NCBI Resources How To Sign in to NCBI

Protein Protein BRAC Search

Create alert Advanced Help

Species
Animals (79,941)
Plants (14,521)
Fungi (12,247)
Protists (4,931)
Bacteria (7,704)
Archaea (34)
Viruses (571)
Customize ...

Source databases
PDB (232)
RefSeq (73,065)
UniProtKB / Swiss-Prot (516)
Customize ...

Genetic compartments
Chloroplast (2)
Mitochondrion (53)
Plasmid (12)
Plastid (2)

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

See **braC** branched-chain amino acid ABC transporter substrate-binding protein **BraC** in the Gene database
braC reference sequences [Protein \(1\)](#)

See the [results of this search \(87 items\)](#) in our new [Identical Protein Groups](#) database.

Items: 1 to 20 of 120705

<< First < Prev Page 1 of 6036 Next > Last >>

[BRAC \[Pseudomonas putida BIRD-1\]](#)

1. **106 aa protein**
Accession: JAO55668.1 GI: 958322777
[BioProject](#) [Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[BraC \[Pseudomonas putida BIRD-1\]](#)

2. **371 aa protein**
Accession: ADR58867.1 GI: 313497501
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#) [Related Sequences](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[branched-chain amino acid transport protein BraC \[Pseudomonas aeruginosa\]](#)

Results by taxon

Top Organisms [\[Tree\]](#)

[Homo sapiens \(1503\)](#)
[Plasmodium falciparum \(1448\)](#)
[Mycobacterium abscessus \(1298\)](#)
[Mus musculus \(1240\)](#)
[Rhizophagus irregularis \(1203\)](#)
[All other taxa \(114013\)](#)
More...

Find related data

Database: [Select](#)

Find items

Search details

BRAC[All Fields]

Search See more...

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-26
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 Gigaxonin
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).

```

/region_name="BTB"
/note="BTB/POZ domain; pfam00651"
/db_xref="CDD:279045"
46..96
/region_name="Domain 2"
/note="NCBI Domains"
49..144
/region_name="BTB"
/note="Broad-Complex, Tramtrack and Bric a brac; smart00225"
/db_xref="CDD:197585"
49..55
/sec_str_type="sheet"
/note="strand 1"
56..62
/sec_str_type="sheet"
/note="strand 2"
63..70
/sec_str_type="helix"
/note="helix 2"
71..79
/sec_str_type="helix"
/note="helix 3"
89..94
/sec_str_type="sheet"
/note="strand 3"
98..109
/sec_str_type="helix"
/note="helix 4"
120..130
/sec_str_type="helix"
/note="helix 5"
133..141
/sec_str_type="helix"
/note="helix 6"

ORIGIN
1 mhhhhhssg vdlgtenlyf qsmavsdpqh aarllralss freesrfcda hlvldgeeip
61 vqknlaaas pyirtklnyn ppkddgstyk ieglegismv mreildyifs gqirlnedti
121 qdvvqaadll lltdktlcc 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/>

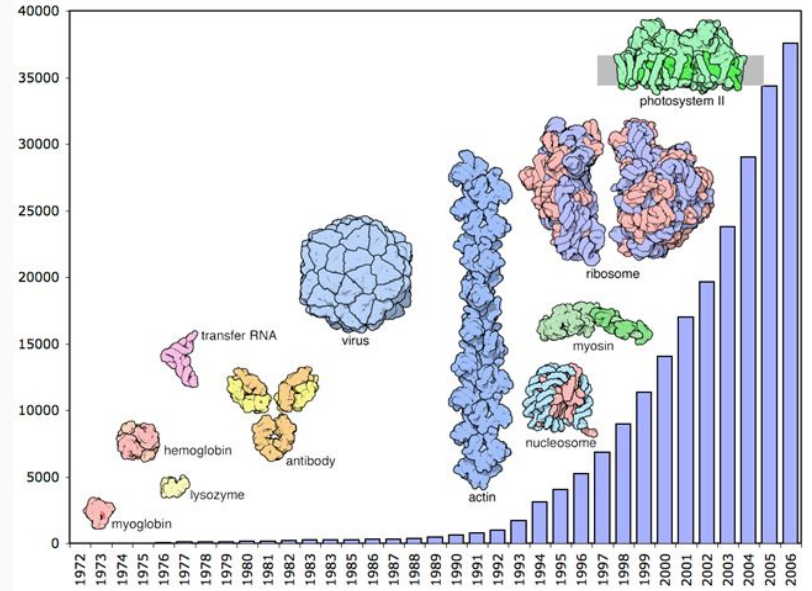
The screenshot shows the UniProt website homepage. At the top, there is a navigation bar with the UniProt logo and a search bar. Below the navigation bar, there are links for "BLAST", "Align", "Retrieve/ID mapping", and "Peptide search". A yellow banner indicates that from June 20, 2018, all traffic will be automatically redirected to HTTPS. The main content area features a grid of database descriptions: UniProt Knowledgebase (Swiss-Prot and TrEMBL), UniRef, UniParc, and Proteomes. A "Supporting data" section is also visible, listing various data types like literature citations, taxonomy, subcellular locations, cross-ref. databases, diseases, and keywords.

UniProtKB/Swiss-Prot entries are tagged with a yellow reviewed icon 

UniProtKB/TrEMBL entries are tagged with a blue unreviewed icon 

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**

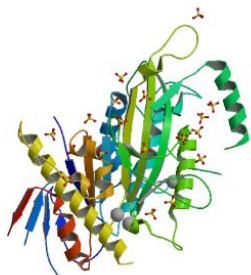
Advanced Search | Browse by Annotations

EMBL-EBI **PDB** PROTEIN DATA BANK | EMBL Data Resource | Wellcome Trust Sanger Institute | PROTEIN DATA BANK | Worldwide Protein Data Bank Foundation

Facebook | Twitter | YouTube | RSS

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

Biological Assembly 1 ?



5X3E
kinesin 6
DOI: [10.2210/pdb5X3E/pdb](https://doi.org/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 | **wwPDB Validation** 3D Report Full Report

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

Metric	Percentile Ranks	Value
Rfree		0.240
Clashscore		7
Ramachandran outliers		0
Sidechain outliers		1.1%
RSRZ outliers		13.0%

Legend: Percentile relative to all X-ray structures; Percentile relative to X-ray structures of similar resolution

This is version 1.0 of the entry. See complete [history](#).

Literature Download Primary Citation

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)

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

Family: *Piwi* (PF02171)

139 architectures 3730 sequences 4 interactions 568 species 103 structures

Summary

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

There are 893 sequences with the following architecture: ArgoN, ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi
[X1WG39_DANRE](#) [Danio rerio (Zebrafish) (Brachydanio rerio)] Uncharacterized protein {ECO:0000313|Ensembl:ENSDFARP00000129194} (858 residues)

[Show](#) all sequences with this architecture.

There are 678 sequences with the following architecture: Piwi
[Z4YLE4_MOUSE](#) [Mus musculus (Mouse)] Piwi-like protein 4 {ECO:0000313|Ensembl:ENSMUSP00000111307} (458 residues)

[Show](#) all sequences with this architecture.

There are 581 sequences with the following architecture: ArgoN, ArgoL1, PAZ, ArgoL2, Piwi
[J3NVY6_GAGT3](#) [Gaeumannomyces graminis var. tritici (strain R3-111a-1) (Wheat and barley take-all root rot fungus)] Uncharacterized protein {ECO:0000313|EMBL:EJT75516.1, ECO:0000313|EnsemblFungi:GGTG_05449T0} (1022 residues)

[Show](#) all sequences with this architecture.

There are 447 sequences with the following architecture: PAZ, Piwi
[V4B7N4_LOTGI](#) [Lottia gigantea (Giant owl limpet)] Uncharacterized protein {ECO:0000313|EMBL:ES084639.1} (791 residues)

[Show](#) all sequences with this architecture.

There are 106 sequences with the following architecture: ArgoL1, PAZ, Piwi
[LSKTH8_PTEAL](#) [Pteropus alecto (Black flying fox)] Piwi-like protein 1 {ECO:0000313|EMBL:ELK14246.1} (821 residues)

[Show](#) all sequences with this architecture.

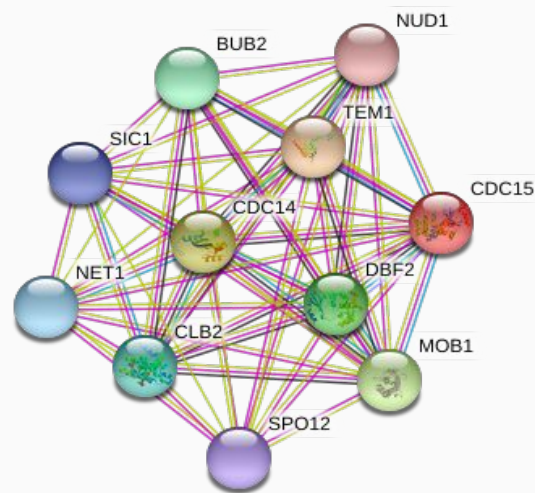
There are 92 sequences with the following architecture: Gly-rich_Ago1, ArgoN, ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi
[V4TXFO_9ROSI](#) [Citrus clementina] Uncharacterized protein {ECO:0000313|EMBL:ESR54581.1} (1036 residues)

[Show](#) all sequences with this architecture.

Jump to...

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
BLAST _N	Nucleotide	Nucleotide
BLAST _P	Protein	Protein
BLAST _X	Protein	Nt. → Protein
TBLAST _N	Nt. → Protein	Protein
TBLAST _X	Nt. → Protein	Nt. → Protein

NAR Database Issue

- Online collection of biological databases:

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

