Biomedical Databases

GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2025-05-07

WEJ LAB

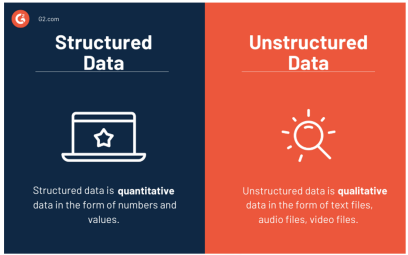
Biological Databases and resources

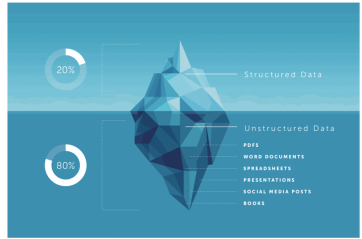
Center for Data Science

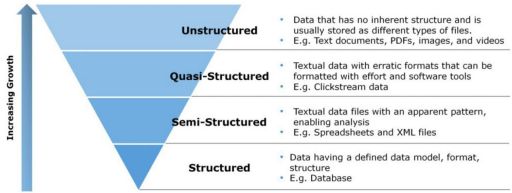
BIG DATA

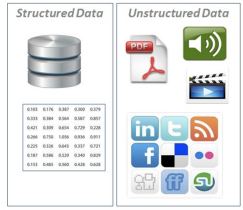
Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data









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

A spreadsheet contains:

- 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

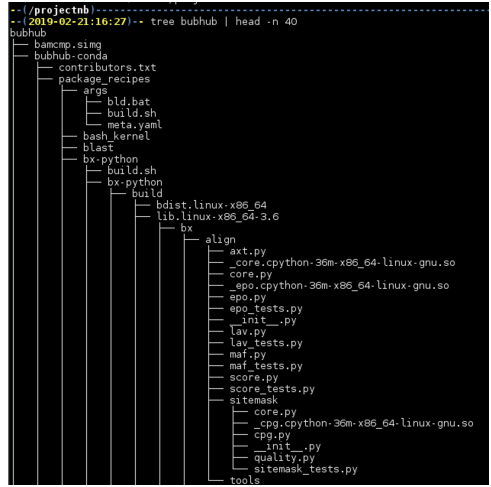
A filesystem can be a Database

Hierarchical data:

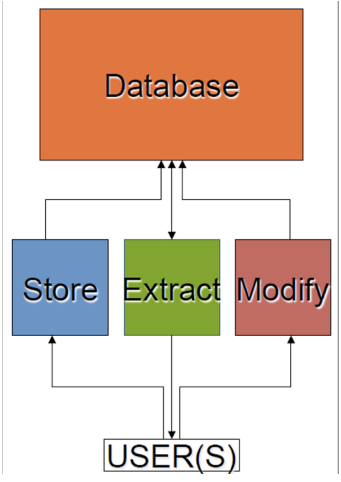
- Some metadata: File, symlink, etc
- Unstructured

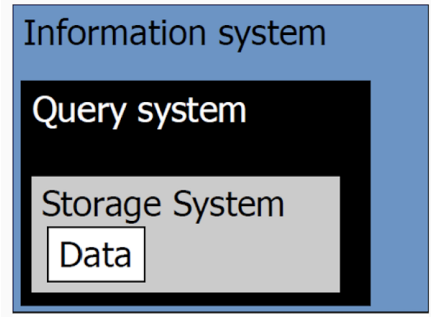
Search tools:

- ls
- find
- grep



Organization and Types of Databases





Organization and Types of Databases

Every database has tools that: Store, Extract, Modify

Flat file databases (flat DBMS)
- Simple, restrictive, table

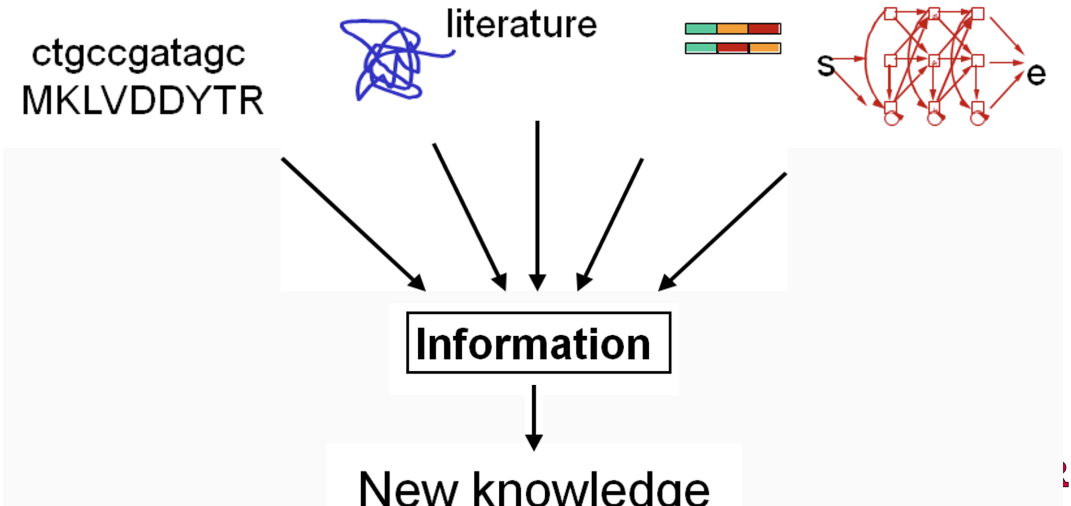
Hierarchical databases
- Simple, restrictive, tables

Relational databases (RDBMS)
- Complex, versatile, tables

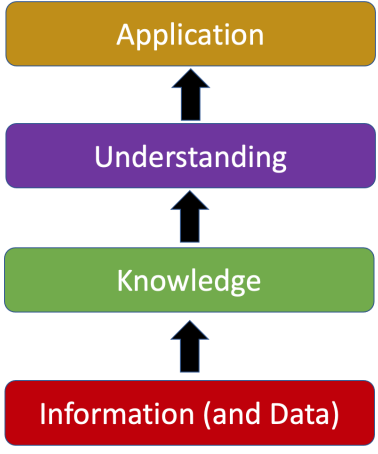
Object-oriented databases (ODBMS)
- Data warehouses and distributed databases

Unstructured databases (object store DBs)

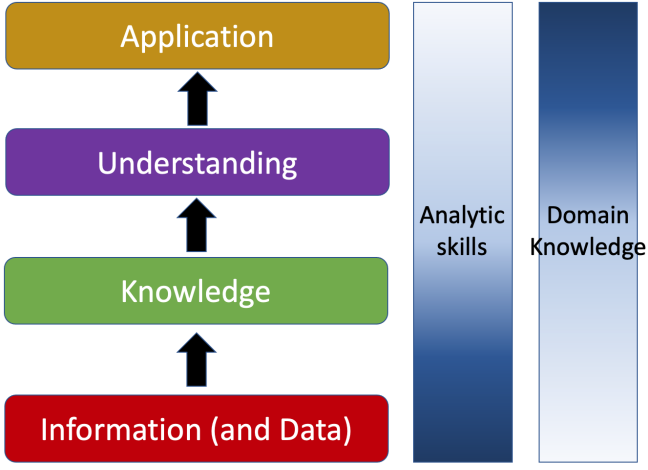
Where do the data come from?



Using biomedical data



Using biomedical data



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:

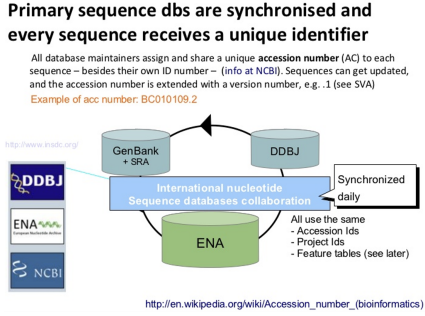
- Results of analysis of primary databases
- Aggregate of many databases
- Content controlled by third party (NCBI)
- Examples: NCBI Protein, Refseq, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain

The NCBI hosts more than 50 different tools and databases

National Center for Biotechnology Information (NCBI):

- **PubMed** - search for citations/papers
- **BLAST** - search for sequences
- **Nucleotide** - all nucleotide sequences (DNA, etc)
- **Genome** - published genomes
- **Protein** - amino acid sequences
- **SNP (dbSNP)** - all human SNPs
- **dbGAP** - Controlled access datasets

International Sequence Database Collaboration



- International Sequence Database Collaboration
- National Centre for Biotechnology Information (NCBI)
- European Nucleotide Archive (ENA)
- DNA Data Bank of Japan (DDBJ)

PubMed is a search engine for literature

- 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

Google Scholar

Google Scholar is another alternative for finding publications:

W. Evan Johnson

FOLLOW

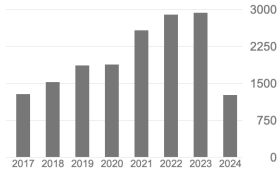
Professor of Medicine
Verified email at rutgers.edu

Data Science Computational Biology Bioinformatics Metagenomics Tuberculosis

TITLE	CITED BY	YEAR
Adjusting batch effects in microarray expression data using empirical Bayes methods WE Johnson, C Li, A Rabinovic Biostatistics 8 (1), 118-127	7202	2007
The sva package for removing batch effects and other unwanted variation in high-throughput experiments JT Leek, WE Johnson, HS Parker, AE Jaffe, JD Storey Bioinformatics 28 (6), 882-883	4502	2012
Tackling the widespread and critical impact of batch effects in high-throughput data JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, ... Nature Reviews Genetics 11 (10), 733-739	2091	2010

Cited by VIEW ALL

	All	Since 2019
Citations	21219	13432
h-index	42	30
i10-index	88	77



Public access VIEW ALL

0 articles 97 articles

WEJ LAB

Collections of DNA/RNA sequences

Center for Data Science

GenBank is an annotated collection of all publicly available DNA sequences

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       MUSNGN      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="MKKLAVVTSRSLPNSPNKYRSFLEWTLNIRYDNKLFISNTNLSRR
                     KLEVTNPIYTRKRSLNIFYLLIPSCRTRLIIWIYIYRNLKRWNTSTVRSRSNSIYRL
                     RPGMKTNIILRCHSYYKPPISNPTVWNNPSWNLAGLLSRQSNLDPILRFPLNLTIYY
                     RCPSNRSFPLPPNRKTQPNRILRCR"
     polyA_site      1803
BASE COUNT     507 a    458 c    311 g    527 t
ORIGIN
        1 tcagttttct ttttttttct ttttttttct ttttttttct ttttttttg ttgattcatg
       61 tcogtttaca tttggtaagt tcacaggoct cagtcaacac aattggactg ctcaggaaat
      121 cctcottggt gacogcagta tacttggoct atgaaccaca gcoacctatg gctaggtagg
      181 agaagctcaa ctgtaggggt gacttggaaa gagaatgcac atggctgtat cgacatttca
      241 catggtggac ctctggccag agtcagcagg ccgaggggttc tcttccgggc tgctcoctca
```

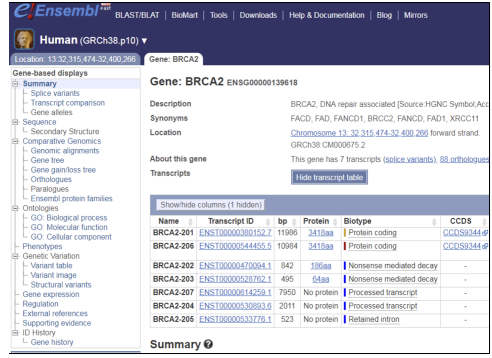
Header

- Title
- Taxonomy
- Citation

Features (AA seq)

Ensembl (European Bioinformatics Institute)

- Comprehensive DNA/RNA sequence and annotation database
- Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data
- Analysis tools:
  - BLAST
  - BioMart
  - Variant Effect Predictor



WEJ LAB

Databases for reference sequences

Center for Data Science

Reference genomes

- NCBI Genome - https://www.ncbi.nlm.nih.gov/datasets/genome/
  - Contains both Genbank and Refseq accessions
- Ensembl - https://useast.ensembl.org/index.html

microRNAs

- miRbase: https://mirbase.org allows you to search and browse miRNAs from a host of species
- Enables download of genomic coordinates for miRNAs (.gff) and sequence (.fa)

rRNAs and tRNAs databases

- rRNAs
- Greengenes - https://greengenes.secondgenome.com
- Silva - https://www.arb-silva.de
- tRNAs
  - gtRNAdb- http://gtrnadb.ucsc.edu

Species-specific databases

For other well studied model organisms, you can often find collections of resources and tools:

- Flybase - D. melanogaster
- Wormbase - C. elegans
- Zfin - D. rerio
- Influenza (link!)

RNACentral

- Developed by EBI and Wellcome Trust
- Integration of many other up-to-date RNA resources and tools

Click here for an overview of core functions

Existing datasets – sequencing data (SRA or GEO)

dbGaP: genotype-phenotype interactions in humans

- Most studies contain PHI and are subject to strict access control and usage regulations
- Applications required to be granted access to datasets
- Any data with sensitive health information must be stored, handled, and interacted with according to appropriate regulations

SRA - Sequencing Read Archive

- Raw sequencing data and alignment info
- Metagenomics, environmental samples, biomedical sequencing

Download via:

- SRA-toolkit
- FTP links on EMBL ENA

A quick demonstration of SRA-toolkit and EMBL-ENA

- SRA-toolkit is the official tool released by NCBI to directly download SRA files
  - Notorious for being obtuse to use and confusing commands / documentation
- EMBL-ENA hosts FTP links directly
  - Most but not all SRA accessions available
  - Have to use wget, curl, or other methods to download

A quick demonstration of SRA-toolkit

Download a file for an asthma host microbiome dataset

```
## attach the sratoolkit
module load sratoolkit

## Save accession to download
acc = "SRR1528344"

## Download using fastq-dump
fastq-dump $acc
# option --split-3 is needed for paired end reads

## don't forget to compress the file!
gzip $acc.fastq
```

A quick demonstration of SRA-toolkit

Download all files for an asthma host microbiome dataset

```
accs=( $( cat SRR_Acc_List.txt ) )
for i in $(seq 0 ${#accs[@]})
do
    fastq-dump ${a[i]};
    gzip ${a[i]}*
done
```

Batch script (SLURM)

```
#!/bin/bash
#SBATCH --job-name=microbiome_download
#SBATCH --mem=1G
#SBATCH --time=01:00:00

module load sratoolkit
cd $HOME/tmp/

accs=( $( cat SRR_Acc_List.txt ) )
for i in $(seq 0 ${#accs[@]})
do
    fastq-dump ${a[i]};
    gzip ${a[i]}*
done
```

Save as a file and use `sbatch` to submit

Batch array (SLURM)

```
#!/bin/bash
#SBATCH --job-name=microbiome_download
#SBATCH --output=asthma.out
#SBATCH --array=0-27
#SBATCH --cpus-per-task=1
#SBATCH --mem=1G
#SBATCH --time=00:20:00

module load sratoolkit
cd /scratch/$USER

accs=( $( cat SRR_Acc_List.txt ) )
acc_number=${accs[$SLURM_ARRAY_TASK_ID]}

fastq-dump --gzip $acc_number
```

Save as a file and use `sbatch` to submit

Gene Expression Data

GEO - Genome Expression Omnibus

- Tied to SRA via Bioproject ID
- GEO also contains processed data, typically specific to a publication
- You may find: Alignments (BAM/SAM), visualizations (.bg, .bed) and intermediate results (delimited formats)

GEOquery example

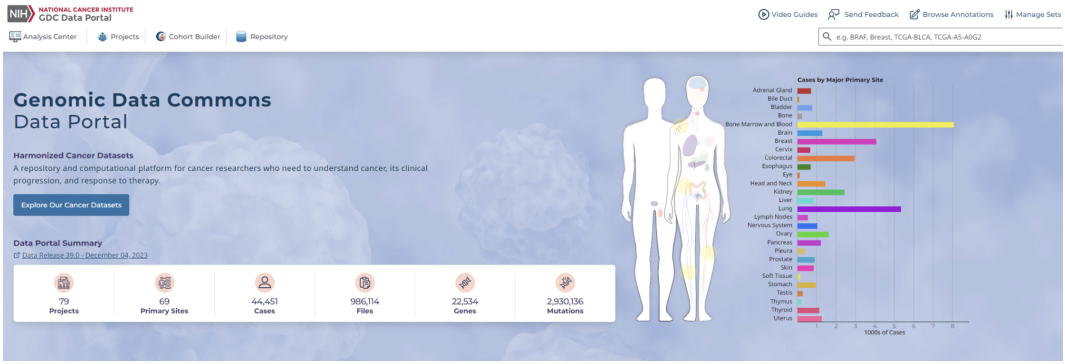
```r
# Load the geoquery library: BiocManager::install("GEOquery")
library(GEOquery)

# Search for a dataset in GEO
geo_search <- getGEO("GSE1297", GSEMatrix = TRUE)

# Display basic information about the dataset
print(geo_search)

# Extract expression data
expression_data <- exprs(geo_search[[1]])

# Display the first few rows of expression data
head(expression_data)
```

WEJ LAB

Domain-specific data

Center for Data Science

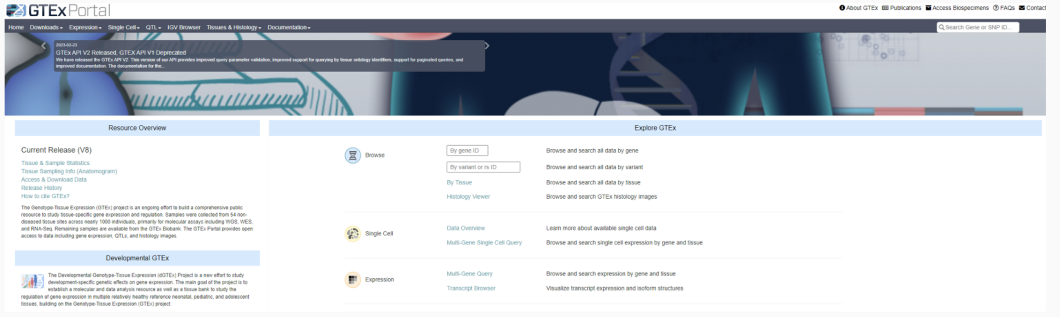
TCGA / NCI Genomic Data Commons



https://portal.gdc.cancer.gov

Tissue-specific gene expression

GTex collects tissue-specific gene expression from primarily healthy individuals



https://www.gtexportal.org/home/

DNA regulatory elements

ENCODE contains a host of information about DNA regulatory elements

https://www.encodeproject.org

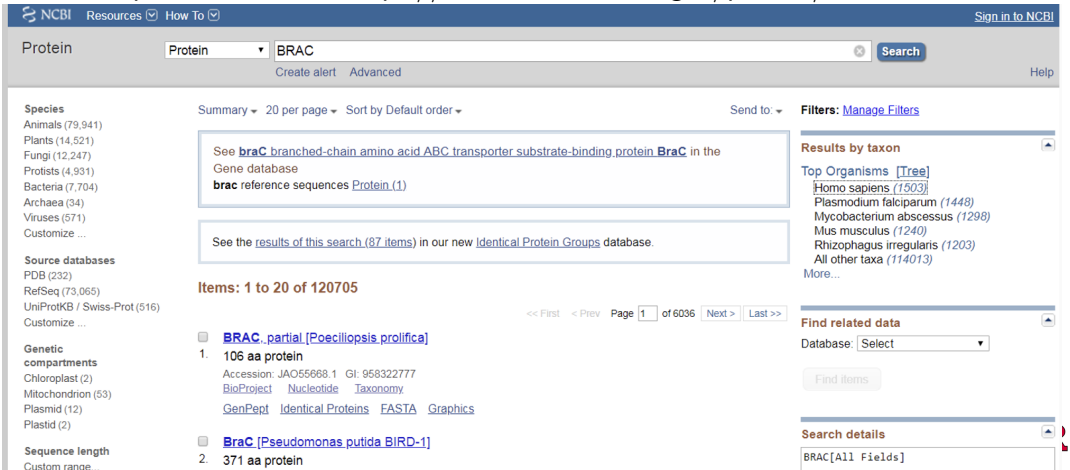
WEJ LAB

Protein databases

Center for Data Science

NCBI Protein

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

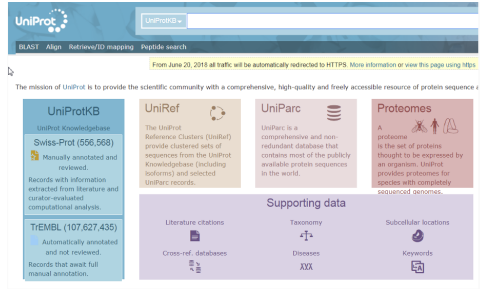
Go to:

```
LOCUS       2PPI_A                144 aa            linear   PRI 26-OCT-2(
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
```

```
Region      /region_name="BTB"
            /note="BTB/POZ domain; pfam00651"
            /db_xref="CDD:279045"
            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"
```

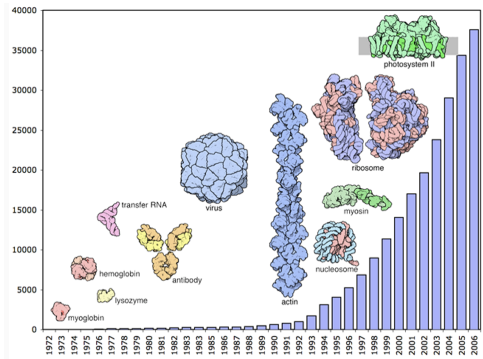
Uniprot

- 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/

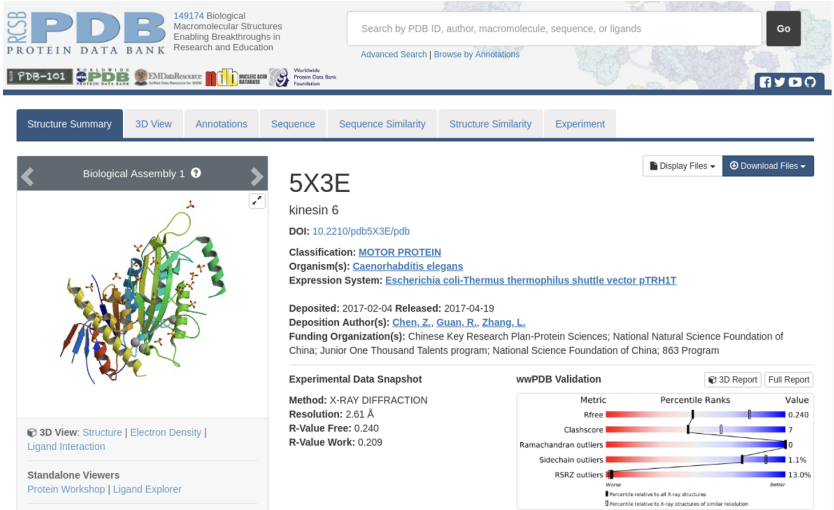


Protein Structure database - PDB

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

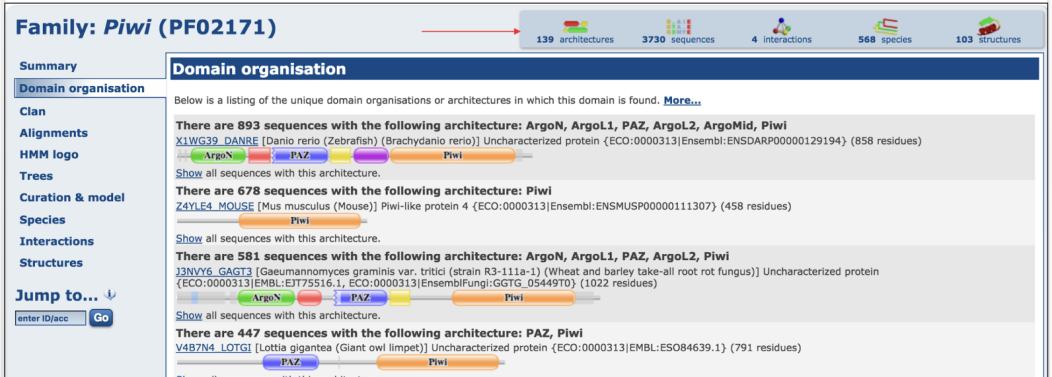


PDB: kinesin 6



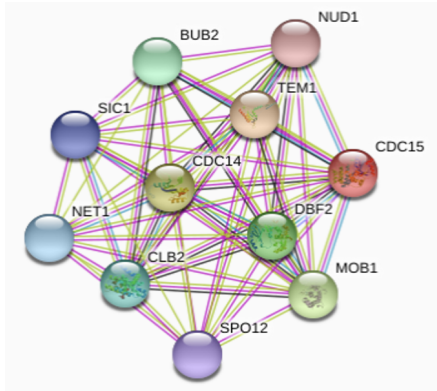
Protein Family Database

- 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

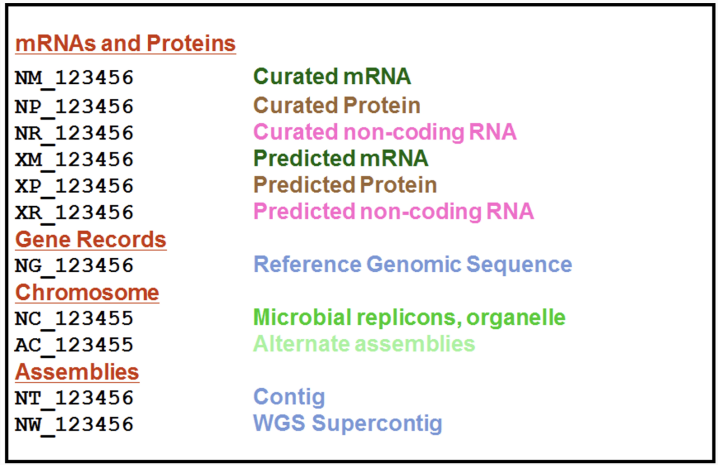


Curated resources

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

RefSeq Annotations



Avoid using outdated or abandoned tools / databases

Keep in mind the following:

- Check when it was last updated (ideally recently)
- Check for an associated peer-reviewed publication
- Check for a github repo
  - How recent was the last commit
  - Are they responsive to issues?

Session info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.4.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.4.2    fastmap_1.2.0     cli_3.6.5         tools_4.4.2
##  [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
##  [9] knitr_1.50        xfun_0.52         digest_0.6.37     rlang_1.1.6
## [13] evaluate_1.0.3
```