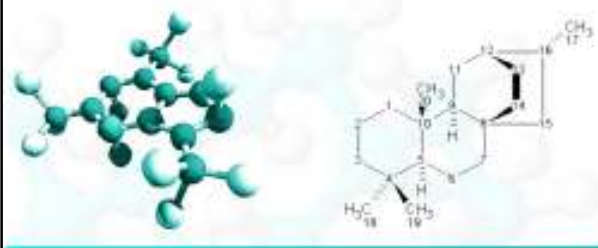


MB11: Bioinformatics Fundamentals

Instructor: Dr. Khalid Raza
 Department of Computer Science,
 Jamia Millia Islamia, New Delhi-110025
kraza@jmi.ac.in Home Page: www.kraza.in



Unit #2

Biological Databases

Contents

- Protein Sequence and Structural Databases
- Nucleotide Sequence Databases
 - EMBL
 - GenBank
 - DDBJ
- Swiss PROT
- PIR
- Protein Data Bank (PDB)
- UniGene
- Saccharomyces Genome Database (SGD)
- PubMed

K. Raza, Jamia Millia Islamia

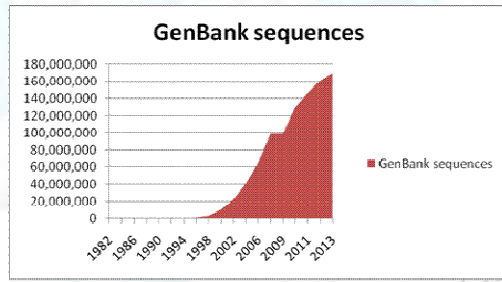
Nucleotide sequence databases

- **GenBank**, **EMBL**, and **DDBJ** are the **three primary nucleotide sequence databases**
- **EMBL** www.ebi.ac.uk/embl/
- **GenBank** www.ncbi.nlm.nih.gov/Genbank/
- **DDBJ** www.ddbj.nig.ac.jp

Genbank

- An **annotated collection of all publicly available nucleotide and proteins**.
- Set up in 1979 at the LANL (Los Alamos).
- Maintained since 1992.
- It is part of the **International Nucleotide Sequence Database Collaboration (INSDC)**.
- INSDC comprises the DNA DataBank of Japan (**DDBJ**), the European Molecular Biology Laboratory (**EMBL**), and **GenBank** at NCBI.
- These three organizations exchange data on a daily basis.

Growth of Genbank sequences



Year	Number of Sequences (Approximate)
1982	0
1986	0
1990	0
1994	0
1998	~10,000,000
2002	~30,000,000
2006	~60,000,000
2009	~100,000,000
2011	~140,000,000
2013	~170,000,000

Access to GenBank

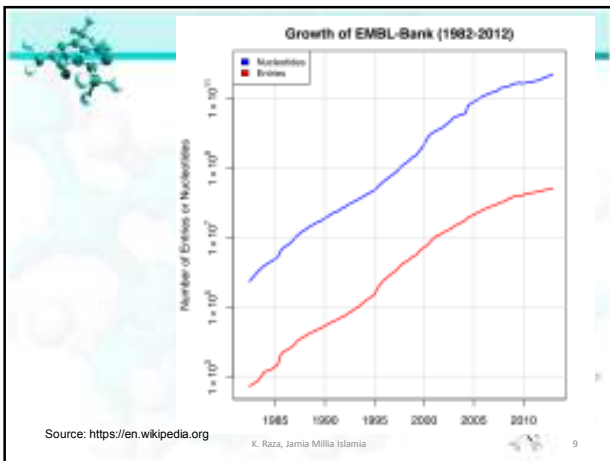
- Several ways to search and retrieve data from GenBank.
- Search for **sequence identifiers and annotations** with **Entrez Nucleotide**.
- It is divided into three divisions:
 - **CoreNucleotide** (the main collection),
 - **dbEST** (Expressed Sequence Tags), and
 - **dbGSS** (Genome Survey Sequences).
- Searching and alignment is done using **BLAST**.
- BLAST searches CoreNucleotide, dbEST, and dbGSS independently.
- Search, link, and download sequences programmatically using **NCBI e-utilities**.

K. Raza, Jamia Millia Islamia

EMBL Nucleotide Sequence Database

- An annotated collection of all publicly available nucleotide and protein sequences.
- Created in 1980 at the *EMBL* in Heidelberg, Germany.
- Maintained since 1994 by EBI- Cambridge.
- Also known as **EMBL-Bank**.
- Now, EMBL Nucleotide Sequence is a section of **Nucleotide Sequence Archive**.
- <http://www.ebi.ac.uk/ena/>
- Increased in size from around 600 entries in 1982 to over **2.5x10⁸** by December 2012.
- Uncompressed file size of **1.6 terabytes**

K. Raza, Jamia Millia Islamia



Accessing ENA content programmatically

These pages describe the ENA Browser REST URL syntax for programmatic users. The majority of ENA data classes and formats are supported by the ENA Browser. Full details about ENA data classes and formats are available here.

Search

Both free text and advanced search can be accessed via REST URLs. These provide access to the complete functionality of ENA's advanced search as well as allowing users to download all data objects that match a given search.

Sequence similarity search

ENAs central NCBI BLAST service can be accessed via REST and SOAP. For assistance matching options to those provided at ENAs sequence search, please

Data retrieval

The main programmatic interface for accessing ENA data is through the ENA Browser. The ENA Browser is designed to be accessed through REST URLs for easy programmatic access to retrieve data and metadata in a variety of formats. The *Taxon* portal also has additional options to allow retrieval of all ENA data based on taxonomic classification.

K. Raza, Jamia Millia Islamia

DNA Data Bank of Japan (DDBJ)

- An annotated collection of all publicly available nucleotide and protein sequences.
- Started, 1984 at the **National Institute of Genetics (NIG)**.
- <http://www.ddbj.nig.ac.jp>

K. Raza, Jamia Millia Islamia

K. Raza, Jamia Millia Islamia

Sequence submission

- Data mainly direct submissions from the authors.
- Submissions through the Internet:
 - Web forms
 - Email
- Sequences shared/exchanged between the 3 centers on a daily basis:
 - The sequence content of the banks is identical.

Sequence Retrieval Tools

- Various tools to get sequences of interests from databases
 - Entrez in NCBI <http://www.ncbi.nlm.nih.gov/Entrez>
 - SRS for EMBL and other DBs <http://srs.ebi.ac.uk>
 - Fetch in GCG package
 - Seqret in EMBOSS

Derived databases

- CUTG Codon usage tabulated from GenBank <http://www.kazusa.or.jp/codon/>
- Genetic Codes Deviations from the standard genetic code in various organisms and organelles <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c>
- UniGene Unified clusters of ESTs and full-length mRNA sequences <http://www.ncbi.nlm.nih.gov/UniGene/>

Protein Databases

- General Sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families

Protein Databases

- **General Sequence databases**
- Protein properties
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families

<http://www.ncbi.nlm.nih.gov/protein>

NCBI Protein database

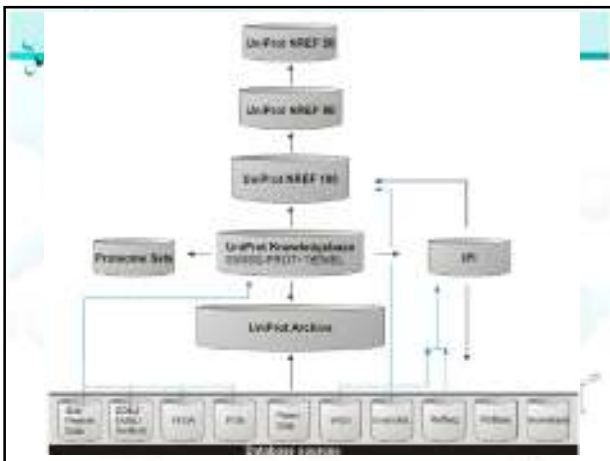
- The NCBI Entrez Protein database are from:
 - SwissProt,
 - Protein Information Resource,
 - Protein Research Foundation,
 - Protein Data Bank, and
 - translations from annotated coding regions in the GenBank and RefSeq databases.

Swiss-Prot

- The Swiss-Prot Protein Knowledgebase is a curated protein sequence database established in 1986.
- Provides a high level of annotation, such as:
 - description of protein function,
 - domains structure,
 - post-translational modifications,
 - variants, etc.
- A minimal level of redundancy and high level of integration with other databases.
- Now, a part of **UniProt**, a "one-stop shop" that allows easy access to all publicly available information of protein sequence annotation.

UniProt

- The **Swiss-Prot**, **TrEMBL**, and **PIR** protein database have united to form the **Universal Protein Resource (UniProt)**
 - UniProt Knowledgebase (UniProtKB): curated sequence information, annotations, linked to other databases.
 - UniProt Reference Clusters (UniRef): removing sequence redundancy by merging sequences that are 100% (UniRef100), 90% (UniRef90) and 50% (UniRef50), linked to Knowledgebase and UniParc records.
 - UniProt Archive (UniParc): history of sequences, no annotation, linked to source records.



UniProt PCNA Search

Search results for PCNA in UniProtKB (accession: P01011) sorted by gene name (ascending).

Accession	Entry name	Status	Protein names	Gene names	Organism	Length
P12004	PCNA_HUMAN	Reviewed	proliferating cell nuclear antigen	PCNA	Homo sapiens (Human)	331
P12004	PCNA_MOUSE	Reviewed	proliferating cell nuclear antigen	PCNA	Mus musculus (Mouse)	331
Q92938	PCNA_CHICK	Reviewed	proliferating cell nuclear antigen	PCNA	Gallus gallus (Chicken)	330
Q92938	PCNA_GALUS	Reviewed	proliferating cell nuclear antigen	PCNA	Gallus gallus (Chicken)	330
P09881	PCNA_PSEU	Reviewed	proliferating cell nuclear antigen	PCNA	Pseudomonas fluorescens (Pseudomonas)	331
P12011	PCNA_DROME	Reviewed	proliferating cell nuclear antigen	PCNA	Drosophila melanogaster (Fruit fly)	330
Q92938	PCNA_THERO	Reviewed	proliferating cell nuclear antigen	pcn (pcn)	Thermoplasma volcanium	331
Q92938	PCNA_ARATH	Reviewed	proliferating cell nuclear antigen 1	PCNA (PCNA1) (PCNA1)	Arabidopsis thaliana (Mustard seed)	330
P12004	PCNA_BOVIN	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Bos taurus (Cattle)	331

UniProt PCNA Search

Search results for PCNA in UniProtKB (accession: P01011) sorted by gene name (ascending).

Accession	Entry name	Status	Protein names	Gene names	Organism	Length
P12011	PCNA_MOUSE	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Mus musculus (Mouse)	331
Q92938	PCNA_CHICK	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Gallus gallus (Chicken)	330
Q92938	PCNA_GALUS	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Gallus gallus (Chicken)	330
P12011	PCNA_DROME	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Drosophila melanogaster (Fruit fly)	330
P12004	PCNA_HUMAN	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Homo sapiens (Human)	331
P12004	PCNA_BOVIN	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Bos taurus (Cattle)	331
P12004	PCNA_PSEU	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Pseudomonas fluorescens (Pseudomonas)	331
P12004	PCNA_THERO	Reviewed	proliferating cell nuclear antigen	pcn (pcn)	Thermoplasma volcanium	331
P12004	PCNA_ARATH	Reviewed	proliferating cell nuclear antigen 1	PCNA (PCNA1) (PCNA1)	Arabidopsis thaliana (Mustard seed)	330
P12004	PCNA_BOVIN	Reviewed	proliferating cell nuclear antigen	PCNA (PCNA)	Bos taurus (Cattle)	331

UniProt PCNA Reviewed Search

Accession	Entry name	Status	Protein names	Date names	Organism	Length
P13085	PCNA_HUMAN	Reviewed	Proliferating cell nuclear antigen	PCNA (P13085) (HUMAN)	Seetherophila	230
P13086	PCNA_MOUSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13086) (MOUSE)	Seetherophila	231
P13087	PCNA_RABBIT	Reviewed	Proliferating cell nuclear antigen	PCNA (P13087) (RABBIT)	Seetherophila	230
P13088	PCNA_CHICK	Reviewed	Proliferating cell nuclear antigen	PCNA (P13088) (CHICK)	Seetherophila	230
P13089	PCNA_SHEEP	Reviewed	Proliferating cell nuclear antigen	PCNA (P13089) (SHEEP)	Seetherophila	230
P13090	PCNA_BOVINE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13090) (BOVINE)	Seetherophila	231
P13091	PCNA_PIG	Reviewed	Proliferating cell nuclear antigen	PCNA (P13091) (PIG)	Seetherophila	231
P13092	PCNA_HORSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13092) (HORSE)	Seetherophila	230
P13093	PCNA_GAL	Reviewed	Proliferating cell nuclear antigen	PCNA (P13093) (GAL)	Seetherophila	230
P13094	PCNA_PHEASANT	Reviewed	Proliferating cell nuclear antigen	PCNA (P13094) (PHEASANT)	Seetherophila	230
P13095	PCNA_CHICKEN	Reviewed	Proliferating cell nuclear antigen	PCNA (P13095) (CHICKEN)	Seetherophila	230
P13096	PCNA_DUCK	Reviewed	Proliferating cell nuclear antigen	PCNA (P13096) (DUCK)	Seetherophila	230
P13097	PCNA_TURKEY	Reviewed	Proliferating cell nuclear antigen	PCNA (P13097) (TURKEY)	Seetherophila	230
P13098	PCNA_GOOSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13098) (GOOSE)	Seetherophila	230
P13099	PCNA_SQUID	Reviewed	Proliferating cell nuclear antigen	PCNA (P13099) (SQUID)	Seetherophila	230

UniProt PCNA BLAST Search

Accession	Entry name	Status	Protein names	Date names	Organism	Length
P13085	PCNA_HUMAN	Reviewed	Proliferating cell nuclear antigen	PCNA (P13085) (HUMAN)	Seetherophila	230
P13086	PCNA_MOUSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13086) (MOUSE)	Seetherophila	231
P13087	PCNA_RABBIT	Reviewed	Proliferating cell nuclear antigen	PCNA (P13087) (RABBIT)	Seetherophila	230
P13088	PCNA_CHICK	Reviewed	Proliferating cell nuclear antigen	PCNA (P13088) (CHICK)	Seetherophila	230
P13089	PCNA_SHEEP	Reviewed	Proliferating cell nuclear antigen	PCNA (P13089) (SHEEP)	Seetherophila	230
P13090	PCNA_BOVINE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13090) (BOVINE)	Seetherophila	231
P13091	PCNA_PIG	Reviewed	Proliferating cell nuclear antigen	PCNA (P13091) (PIG)	Seetherophila	231
P13092	PCNA_HORSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13092) (HORSE)	Seetherophila	230
P13093	PCNA_GAL	Reviewed	Proliferating cell nuclear antigen	PCNA (P13093) (GAL)	Seetherophila	230
P13094	PCNA_PHEASANT	Reviewed	Proliferating cell nuclear antigen	PCNA (P13094) (PHEASANT)	Seetherophila	230
P13095	PCNA_CHICKEN	Reviewed	Proliferating cell nuclear antigen	PCNA (P13095) (CHICKEN)	Seetherophila	230
P13096	PCNA_DUCK	Reviewed	Proliferating cell nuclear antigen	PCNA (P13096) (DUCK)	Seetherophila	230
P13097	PCNA_TURKEY	Reviewed	Proliferating cell nuclear antigen	PCNA (P13097) (TURKEY)	Seetherophila	230
P13098	PCNA_GOOSE	Reviewed	Proliferating cell nuclear antigen	PCNA (P13098) (GOOSE)	Seetherophila	230
P13099	PCNA_SQUID	Reviewed	Proliferating cell nuclear antigen	PCNA (P13099) (SQUID)	Seetherophila	230

UniProt PCNA Alignment

Accession	Entry name	Protein names	Date names	Organism	Length
P13085	PCNA_HUMAN	Proliferating cell nuclear antigen (PCNA) (Human)	PCNA (P13085) (HUMAN)	Seetherophila	230
P13086	PCNA_MOUSE	Proliferating cell nuclear antigen (PCNA) (Mouse)	PCNA (P13086) (MOUSE)	Seetherophila	231
P13087	PCNA_RABBIT	Proliferating cell nuclear antigen (PCNA) (Rabbit)	PCNA (P13087) (RABBIT)	Seetherophila	230
P13088	PCNA_CHICK	Proliferating cell nuclear antigen (PCNA) (Chicken)	PCNA (P13088) (CHICKEN)	Seetherophila	230
P13089	PCNA_SHEEP	Proliferating cell nuclear antigen (PCNA) (Sheep)	PCNA (P13089) (SHEEP)	Seetherophila	230
P13090	PCNA_BOVINE	Proliferating cell nuclear antigen (PCNA) (Bovine)	PCNA (P13090) (BOVINE)	Seetherophila	231
P13091	PCNA_PIG	Proliferating cell nuclear antigen (PCNA) (Pig)	PCNA (P13091) (PIG)	Seetherophila	231
P13092	PCNA_HORSE	Proliferating cell nuclear antigen (PCNA) (Horse)	PCNA (P13092) (HORSE)	Seetherophila	230
P13093	PCNA_GAL	Proliferating cell nuclear antigen (PCNA) (Gal)	PCNA (P13093) (GAL)	Seetherophila	230
P13094	PCNA_PHEASANT	Proliferating cell nuclear antigen (PCNA) (Pheasant)	PCNA (P13094) (PHEASANT)	Seetherophila	230
P13095	PCNA_CHICKEN	Proliferating cell nuclear antigen (PCNA) (Chicken)	PCNA (P13095) (CHICKEN)	Seetherophila	230
P13096	PCNA_DUCK	Proliferating cell nuclear antigen (PCNA) (Duck)	PCNA (P13096) (DUCK)	Seetherophila	230
P13097	PCNA_TURKEY	Proliferating cell nuclear antigen (PCNA) (Turkey)	PCNA (P13097) (TURKEY)	Seetherophila	230
P13098	PCNA_GOOSE	Proliferating cell nuclear antigen (PCNA) (Goose)	PCNA (P13098) (GOOSE)	Seetherophila	230
P13099	PCNA_SQUID	Proliferating cell nuclear antigen (PCNA) (Squid)	PCNA (P13099) (SQUID)	Seetherophila	230

Protein Databases

- General Sequence databases
- **Protein properties**
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families

DBs based on Protein properties

- **AAindex**: A database of amino acid indices and amino acid mutation matrices.
- **Cybase**: Cyclic proteins
- **dbPTM**: protein post-translational modification (PTM) information
- **iProLINK**: Integrated Protein Literature, INformation and Knowledge
- **PFD** - Protein Folding Database
- **PINT**: Protein-protein Interactions Thermodynamic Database
- **MIPS, STRING**: Protein-protein Interactions Database
- **REFOLD**: Data related to refolding experiments

Protein sequence motifs and active sites

- **ASC** - Active Sequence Collection
- **Blocks**
- **COMe** - Co-Ordination of Metals etc.
- **CSA** - Catalytic Site Atlas
- **eBLOCKS**
- **eMOTIF**
- **InterPro**
- **Metalloprotein Site Database**
- **O-GLYCBASE**
- **PDBSite**
- **PhosphoELM Base**
- **PRINTS**
- **PROMISE**
- **ProRule**
- **PROSITE**

Protein domain databases; protein classification

- ADDA - Automatic Domain Decomposition Algorithm
- BALIBASE
- BIOZON
- CluStr - Clusters of Swiss-Prot and TrEMBL proteins
- COG - Clusters of Orthologous Groups of proteins
- FusionDB
- HSSP
- InterDom
- InterPro <---- PROSITE, Pfam, PRINTS, Prodom, SMART, TIGRFAMS, PIR superfamily
- iProClass
- MulPSSM
- PALI
- Pfam
- ProDom <---- SP, TrEMBL
- SUPFAM

Databases of individual protein families

- AARSDb
- ABCdb
- ARAMEMNON
- BacTregulators (formerly AraC/XylS database)
- CSDb - Cold Shock Domain database
- DCCP - Database of Copper-Chelating Proteins
- DExH/D Family Database
- DSD
- Endogenous GPCR List
- EROP-Moscow
- ESTHER
- FUNPEP
- GPCRDB
- gpDB - G-protein database
- Histone Database
- HIV RT and Protease Sequence Database
- Homeobox Page
- Homeodomain Resource
- InBase
- KinG - Kinases in Genomes
- Knottins
- LGICdb
- Lipase Engineering Database
- Lipid MAPS
- MEROPS
- Nuclear Receptor Resource
- NucleaRDB
- NUREBASE
- PHYTOPROT
- Protein kinase resource
- Ribonuclease P Database
- RNRdb
- RTKdb - Receptor Tyrosine Kinase database

Database Searching Tips

- Look for links to **Help** or **Examples**
- Always check update dates
- Level of curation
- Try **Boolean** searches
- Be careful with UK/US **spelling** differences
 - leukaemia vs leukemia
 - haemoglobin vs hemoglobin
 - colour vs color

Protein Structures

Tertiary protein structure prediction is possibly the Holy Grail of bioinformatics.

This houses a collection of 3D coordinates of each atom in a protein, allowing the structure to be displayed by viewing software.

Protein structures are submitted by individual researchers and have been determined by x-ray diffraction, or NMR.

K. Raza, Jamia Millia Islamia 34

Protein Structure Databases

- **PDB:** Protein DataBank, New Jersey, USA
 - <http://www.rcsb.org/>
- **EMSD:** EBI Macromolecular Structure Database
 - Management and distribution of data on macromolecular structures in close collaboration with the PDB.
 - <http://www.ebi.ac.uk/msd/index.html>
- **SCOP:** Structural Classification of Proteins
 - scop.mrc-lmb.cam.ac.uk/scop
- **CATH:** Classification, Architecture, Topology, Homology
 - http://www.biochem.ucl.ac.uk/bsm/cath_new/

K. Raza, Jamia Millia Islamia 35

Selected NMR database resources for macromolecular structures.

Database	URL
Structure and sequence/structure databases	
SCOP	http://scop.nyu.edu/
CATH	http://www.ebi.ac.uk/cath/
PROSP	http://www2.chem.ucl.ac.uk/prosp/
Molecular Modeling Database	http://www.molmod.org/
CSMPDB	http://www.cysdb.joe.cam.ac.uk/cysmpdb/
CSO	http://www.unimelb.edu.au/csdb/
Library of Protein Family Cross (LPCF)	http://www.rcsb.org/ligand/EDU/projects/bsm/LPCF/
3D_AU (a database of aligned protein structures and related sequences)	http://www.mol.fhb.org/delegated/3d_au.html
EDRF (a relational database and query tool for proteins)	http://www.chem.ucl.ac.uk/edrf/
HSSP	http://www.scripps.edu/hssp/
Specialty databases	
HIV Protein Database	http://www.rcsb.org/hiv/
Nucleic Acid Database	http://www.rcsb.org/nucleic/
Protein Synthesis and Protease Inhibitor (PSI) server	http://www.rcsb.org/protein/synthesis/
International Nucleic Acid Database (INCD)	http://www.rcsb.org/inca/
Diagrams Database	http://www.rcsb.org/diagrams/
Features Database	
Molecular Movements Database	http://www.rcsb.org/movements/
CLUSTAL	http://www.rcsb.org/clustal/
PROSP	http://www.rcsb.org/prosp/
Protein Quaternary Structures (PQS)	http://www.rcsb.org/pqs/
Structure (Protein-Signal complex database)	http://www.rcsb.org/structure/
PROSP	http://www.rcsb.org/prosp/
PROSP	http://www.rcsb.org/prosp/
PROSP	http://www.rcsb.org/prosp/
Biological Macromolecule Crystallography Database (EMSD)	http://www.rcsb.org/emsd/
Resources	
Protein Data Bank	http://www.rcsb.org/pdb/

Protein Data Bank (PDB)

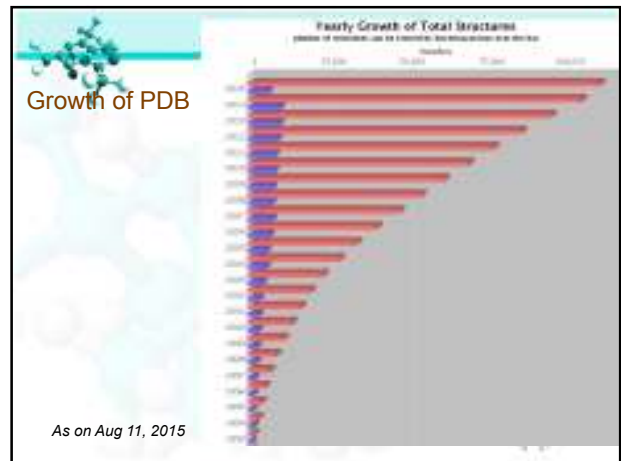
- A repository for 3-D biological macromolecular structure.
- It includes **proteins**, **nucleic acids** and **viruses**.
- Obtained by X-Ray crystallography (80%) or NMR spectroscopy (16%).
- Important in solving real problems in molecular biology.
- Established in 1972 at Brookhaven National Laboratory.
- Transferred to the **Research Collaboratory for Structural Bioinformatics (RCSB)** in 1998.
- Sole international repository of macromolecular structure data.
- URL: <http://www.rcsb.org/>

Protein Data Bank (Structures)

PDB Contents

PDB Current Holdings Breakdown

Exp. Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	92817	1649	4616	4	99086
NMR	9699	1126	227	8	11060
ELECTRON MICROSCOPY	602	29	199	0	830
HYBRID	70	3	2	1	76
other	166	4	6	13	189
Total	103354	2811	5050	26	1,11,241



Effective use of PDB

- Queries are of three types
 - PDB id - As quoted in paper
 - Search Lite - one or more keywords
 - Search Fields - A detailed query form
- Query results
 - Structure Explorer - details of the structure
 - Query Result Browser - for multiple structures
- PDB Viewer

Structure Explorer

PDB data formats

- PDB file format was used to contain the coordinates and related information.
- In the late 1990's, macromolecular Crystallographic Information file (**mmCIF**) evolved.
- mmCIF and PDBML
 - Push in to make structure files completely self-contained descriptions of the experiment and details of the structure determination.
 - **PDB file format unstructured and obsolete**

PDB File Format

- Text file – you can edit with a text editor e.g. WordPad
- Atomic co-ordinates
- Rich annotation
 - Citation
 - Experimental Method
 - Biological source e.
 - Etc.

PDB: Example

- HEADER
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SECRES
- ATOM COORDINATES

```

HEADER  LYASE(OXO-ACID) 01-OCT-91 12CA 12CA 2
COMPND  CARBONIC ANHYDRASE /II (CARBONATE DEHYDRATASE) (HCA II) 12CA 3
COMPND  2 (E.C.4.2.1.1) MUTANT WITH VAL 121 REPLACED BY ALA (V121A) 12CA 4
SOURCE  HUMAN (HOMO SAPIENS) RECOMBINANT PROTEIN 12CA 5
AUTHOR  S.K.NAIR,D.W.CHRISTIANSON 12CA 6
REVDAT  1 15-OCT-92 12CA 0 12CA 7
JRNL  AUTH  S.K.NAIR,T.L.CALDERONE,D.W.CHRISTIANSON,C.A.FIERKE 12CA 8
JRNL  TITL  ALTERING THE MOUTH OF A HYDROPHOBIC POCKET. 12CA 9
JRNL  TITL  2 STRUCTURE AND KINETICS OF HUMAN CARBONIC ANHYDRASE 12CA 10
JRNL  TITL  3 /II MUTANTS AT RESIDUE VAL-121 12CA 11
JRNL  REF  J.BIOL.CHEM. V. 266 17320 1991 12CA 12
JRNL  REFN  ASTM JBCHA3 US ISSN 0021-9258 071 12CA 13
REMARK  1 12CA 14
REMARK  2 RESOLUTION. 2.4 ANGSTROMS. 12CA 15
REMARK  2 RESOLUTION. 2.4 ANGSTROMS. 12CA 16
REMARK  3 12CA 17
REMARK  3 REFINEMENT. 12CA 18
REMARK  3 PROGRAM PROLSQ 12CA 19
REMARK  3 AUTHORS HENDRICKSON,KONNERT 12CA 20
REMARK  3 R VALUE 0.170 12CA 21
REMARK  3 RMSD BOND DISTANCES 0.011 ANGSTROMS 12CA 22
REMARK  3 RMSD BOND ANGLES 1.3 DEGREES 12CA 23
REMARK  4 12CA 24
REMARK  4 N-TERMINAL RESIDUES SER 2, HIS 3, HIS 4 AND C-TERMINAL 12CA 25
REMARK  4 RESIDUE LYS 260 WERE NOT LOCATED IN THE DENSITY MAPS AND, 12CA 26
REMARK  4 THEREFORE, NO COORDINATES ARE INCLUDED FOR THESE RESIDUES. 12CA 27
            
```

PDB (cont.)

```

SHEET  3 S10 PHE 66 PHE 70-1 O ASN 67 N LEU 60 12CA 68
SHEET  4 S10 TYR 88 TRP 97-1 O PHE 93 N VAL 68 12CA 69
SHEET  5 S10 ALA 116 ASN 124-1 O HIS 119 N HIS 94 12CA 70
SHEET  6 S10 LEU 141 VAL 150-1 O LEU 144 N LEU 120 12CA 71
SHEET  7 S10 VAL 207 LEU 212 1 O ILE 210 N GLY 145 12CA 72
SHEET  8 S10 TYR 191 GLY 196-1 O TRP 192 N VAL 211 12CA 73
SHEET  9 S10 LYS 257 ALA 258-1 O LYS 257 N THR 193 12CA 74
SHEET 10 S10 LYS 39 TYR 40 1 O LYS 39 N ALA 258 12CA 75
TURN  1 TI GLN 28 VAL 31 TYPE VIB (CIS-PRO 30) 12CA 76
TURN  2 I2 GLY 81 LEU 84 TYPE II(PRIME) (GLY 82) 12CA 77
TURN  3 T3 ALA 134 GLN 137 TYPE I (GLN 136) 12CA 78
TURN  4 T4 GLN 137 GLY 140 TYPE I (ASP 139) 12CA 79
TURN  5 T5 THR 200 LEU 203 TYPE VIA (CIS-PRO 202) 12CA 80
TURN  6 T6 GLY 233 GLU 236 TYPE II (GLY 235) 12CA 81
CRYST1 42.700 41.700 73.000 90.00 104.60 90.00 P 21 2 12CA 82
ORIGX1 1.000000 0.000000 0.000000 0.00000 12CA 83
ORIGX2 0.000000 1.000000 0.000000 0.00000 12CA 84
ORIGX3 0.000000 0.000000 1.000000 0.00000 12CA 85
SCALE1 0.023419 0.000000 0.006100 0.00000 12CA 86
SCALE2 0.000000 0.023981 0.000000 0.00000 12CA 87
SCALE3 0.000000 0.000000 0.014156 0.00000 12CA 88
ATOM  1 N TRP 5 8.519 -0.751 10.738 1.00 13.37 12CA 89
ATOM  2 CA TRP 5 7.743 -1.668 11.585 1.00 13.42 12CA 90
ATOM  3 C TRP 5 6.786 -2.502 10.667 1.00 13.47 12CA 91
ATOM  4 O TRP 5 6.422 -2.085 9.607 1.00 13.57 12CA 92
ATOM  5 CB TRP 5 6.997 -0.917 12.645 1.00 13.34 12CA 93
ATOM  6 CG TRP 5 5.784 -0.209 12.221 1.00 13.40 12CA 94
ATOM  7 CD1 TRP 5 5.681 1.084 11.797 1.00 13.29 12CA 95
            
```


Fragment of CIF example

```

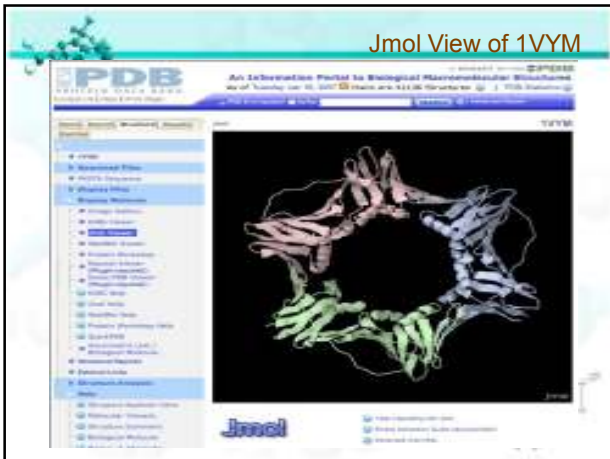
#####
# ATOM_SITE #
#####
loop_
  _atom_site.label_seq_id
  _atom_site.group_PDB
  _atom_site.type_symbol
  _atom_site.label_atom_id
  _atom_site.label_comp_id
  _atom_site.label_asym_id
  _atom_site.auth_seq_id
  _atom_site.label_alt_id
  _atom_site.cartn_x
  _atom_site.cartn_y
  _atom_site.cartn_z
  _atom_site.occupancy
  _atom_site.B_iso_or_equiv
  _atom_site.footer_id
  _atom_site.label_entity_id
  _atom_site.id
1
ATOM N N GLY A 1 . -8.863 16.944 14.289 1.00 21.88 1 1
1
ATOM C CA GLY A 1 . -9.929 17.026 13.244 1.00 22.85 1 1
2
ATOM C C GLY A 1 . -10.051 15.625 12.618 1.00 43.92 1 1
3
1
ATOM O O GLY A 1 . -9.782 14.728 13.407 1.00 25.22 1 1
            
```

Structure Visualization

PDB file =====>
Rasmol/Pymol/Jmol



K. Raza, Jamia Millia Islamia 48



3-D Structure from PDB

- 20 Amino acids

<http://www.clunet.edu/BioDev/omm/aa/aa.htm>
<http://www.nyu.edu/pages/mathmol/library/life/>
http://inquiry.uiuc.edu/bioweb/tutorial/amino_acids.htm

How to Construct 3-D Molecule

- Read coordinates from PDB.
- Set up data structure of molecules
- Form bonds among atoms and groups
- Calculate secondary structure.
- Implement 3-D graphical algorithms.
- Render 3-D graph in various style, wires, sticks, balls, ribbons, and the like.

Bonds among atoms


ATOM	20	N	LEU	1	4	30.279	-25.716	105.041	1.00	10.60	2MCG 249
ATOM	21	CA	LEU	1	4	31.406	-26.518	104.496	1.00	9.39	2MCG 250
ATOM	22	C	LEU	1	4	32.659	-25.786	105.165	1.00	8.90	2MCG 251
ATOM	23	O	LEU	1	4	32.890	-24.586	104.967	1.00	8.74	2MCG 252
ATOM	24	CB	LEU	1	4	31.615	-26.794	103.141	1.00	8.79	2MCG 253
ATOM	25	CG	LEU	1	4	31.552	-27.440	101.860	1.00	8.37	2MCG 254
ATOM	26	CD1	LEU	1	4	32.732	-26.945	100.970	1.00	7.99	2MCG 255
ATOM	27	CD2	LEU	1	4	31.706	-28.963	102.016	1.00	8.09	2MCG 256

$$\begin{array}{c}
 \text{COO}^- \\
 | \\
 \text{H}-\text{C}-\text{CH}_2-\text{CH} \\
 | \quad \quad | \\
 \text{NH}_3^+ \quad \quad \text{CH}_3 \\
 \quad \quad \quad | \\
 \quad \quad \quad \text{CH}_3
 \end{array}$$

Leucine LEU L
Formula: $\text{C}_6\text{H}_{13}\text{NO}_2$



Bonds between groups

ATOM	9	N	SER	1	2	25.548	-22.930	103.333	1.00	16.05	2MCG 238
ATOM	10	CA	SER	1	2	26.608	-22.758	104.327	1.00	15.38	2MCG 239
ATOM	11	C	SER	1	2	27.351	-24.076	104.604	1.00	14.81	2MCG 240
ATOM	12	O	SER	1	2	27.530	-24.949	103.740	1.00	15.00	2MCG 241
ATOM	13	CB	SER	1	2	25.887	-22.406	105.682	1.00	15.73	2MCG 242
ATOM	14	OG	SER	1	2	25.193	-23.586	106.117	1.00	15.14	2MCG 243
ATOM	15	N	ALA	1	3	27.758	-24.228	105.876	1.00	13.72	2MCG 244
ATOM	16	CA	ALA	1	3	28.328	-25.397	106.456	1.00	12.33	2MCG 245
ATOM	17	C	ALA	1	3	29.255	-26.303	105.686	1.00	11.58	2MCG 246
ATOM	18	O	ALA	1	3	29.033	-27.552	105.641	1.00	11.28	2MCG 247
ATOM	19	CB	ALA	1	3	27.101	-26.228	106.998	1.00	12.39	2MCG 248
ATOM	20	N	LEU	1	4	30.279	-25.716	105.041	1.00	10.60	2MCG 249
ATOM	21	CA	LEU	1	4	31.406	-26.518	104.496	1.00	9.39	2MCG 250
ATOM	22	C	LEU	1	4	32.658	-25.786	105.165	1.00	8.90	2MCG 251
ATOM	23	O	LEU	1	4	32.890	-24.586	104.967	1.00	8.74	2MCG 252
ATOM	24	CB	LEU	1	4	31.615	-26.794	103.141	1.00	8.79	2MCG 253
ATOM	25	CG	LEU	1	4	31.552	-27.440	101.860	1.00	8.37	2MCG 254
ATOM	26	CD1	LEU	1	4	32.732	-26.945	100.970	1.00	7.99	2MCG 255
ATOM	27	CD2	LEU	1	4	31.706	-28.963	102.016	1.00	8.09	2MCG 256



Nucleic Acid Database (NDB)



- NDB contains information about experimentally-determined nucleic acids and complex assemblies.
- Use the NDB to perform searches based on annotations relating to sequence, structure and function, and to download, analyze, and learn about nucleic acids.
- The NDB Project is funded by the National Science Foundation and the Department of Energy.
- The goal of NDBP is to assemble and distribute structural information about nucleic acids,
- The format of NDB is the same as PDB.

Nucleic Acid Database (NDB)

A Portal for Three-dimensional Structural Information about Nucleic Acids.



As of 12-Aug-2015 number of released structures:
7671

UniGene

- **UniGene** is an NCBI database of the **transcriptome**.
- Information on protein similarities, gene expression, cDNA clones, and genomic location is included with each entry.


K. Raza, Jamia Millia Islamia 57

Saccharomyces Genome Database (SGD)


- Provides Internet access to the complete *Saccharomyces cerevisiae* (yeast) genomic sequence, its genes and their products, the phenotypes of its mutants, and the literature supporting these data.
- **Currently the only complete sequence of a eukaryotic genome.**
- Aids researchers by providing basic information and tools for sequence similarity searching, and finding relationships between genes.
- SGD presents information using a variety of user-friendly, dynamically created graphical displays illustrating physical, genetic and sequence feature maps.
- <http://www.yeastgenome.org/>

K. Raza, Jamia Millia Islamia 58





PubMed

- Where we search for research from medical journals
- 13 million citations/abstracts
- 4,500 journals
 - From 1966
- Covers
 - Medicine
 - Nursing
 - Dentistry
 - Veterinary medicine
 - Healthcare systems
 - Preclinical sciences

More About PubMed

- Links to other databases
 - OLDMEDLINE (1950-1965)
 - International biomedical journals
 - 1,760,000 citations
- Not all journals are strictly scientific or medical
- Links (LinkOut) to the full-text of articles at participating publishers' web sites



What Can You Do in PubMed?

- Search for articles (usually abstracts)
 - By keyword
 - By author
 - By Journal, etc
- Combine searches
- Link to related articles
- Link to outside sources
 - To purchase the full article
 - Look at related books (including pages in the books)
- Clinical queries

The PubMed Homepage



<http://www.ncbi.nlm.nih.gov/pubmed>

PubMed – Search Overview

- Search by keyword
 - E.g., thrombin, but you get 31,365 results
- Use a more specific keyword
 - Like “Thrombin-JMI”, but now you only get 2 hits
- Additional tools
 - Limits
 - Search Field Tags
 - meSH (Medical Subject Heading)

PubMed – Additional Tools

- Limits
 - Age, groups, gender, type of study, etc
- Search Field tags
 - Qualify terms, used in brackets [ta]
- Boolean Statements
 - And, OR, NOT
- meSH
 - The “vocabulary” of PubMed
 - Stands for Medical Subjects Headings
 - Used for indexing biomedical literature

PubMed -- Limits



PubMed -- Search Field Tags

- Search Field Tags also qualify terms
- Letters in brackets
 - Nature [ta] (title abbreviation)
 - Watson JD [au] (author)
 - 1953 [dp] (date of publication)
 - DNA [mh] (meSH)

Boolean Statements

- Example: citations on DNA authored by Crick in 1993
– Dna [mh] AND crick [au] AND 1993[dp]
- Effect of heat or humidity on surgical hemostasis
– (heat OR humidity) AND surgical hemostasis

PubMed -- The meSH Tree

- meSH organizes terms in a hierarchal manner



- Information on over 650 diseases and conditions
- Medical encyclopedia and dictionary
- Information on prescription and nonprescription drugs
- Links to ClinicalTrials.gov
- Links to news
- Sponsored by the NIH.

<http://www.nlm.nih.gov/medlineplus/>

References

- Rastogi et al., *Bioinformatics: Methods & Applications*, PHI.
- Jin Xiong, *Essential Bioinformatics*, Cambridge University Press.
- Several Online Resources....

What's Next?

- **Data Analysis Tools**
 - DNA Sequence Analysis Tools
 - RNA analysis tools
 - Protein sequence and structure analysis tools
 - Microarray data analysis tools
 - Literature search tools