

Unit-2

Introduction to Biological data bases

→ Biological Database.

Bio-life, Database - collection of data & arranged & managed

→ Biological Database is the a collection of biological information that is organized so that it can be easily accessed, managed & updated.

→ Data :-

① DNA sequence of gene or full genome.

② sequence of protein and 3-D structure of protein, nucleic acids and protein-nucleic acid complex, transcriptomic

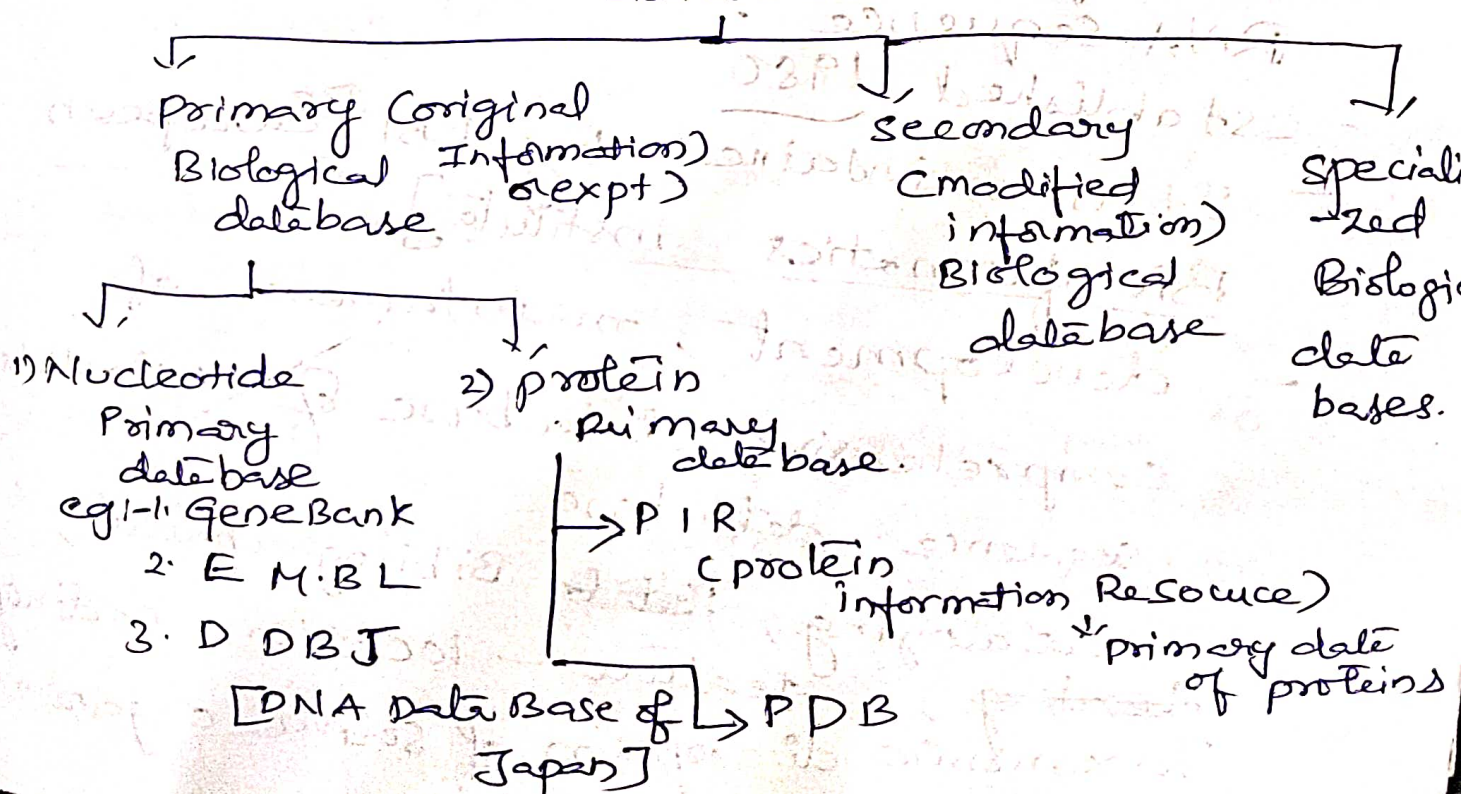
- There are 15,52 database - publically accessible

- Genomics, proteins.

- Bibliographic databases - Literature

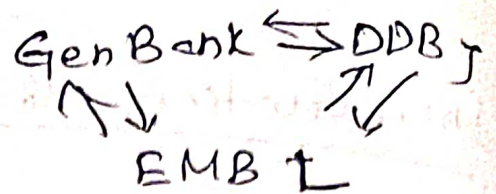
- Taxonomic databases - classification

Biological data bases



Gene Bank

- Nucleotide data
- Primary data
- It has Flat file structure.



↓
It access text file

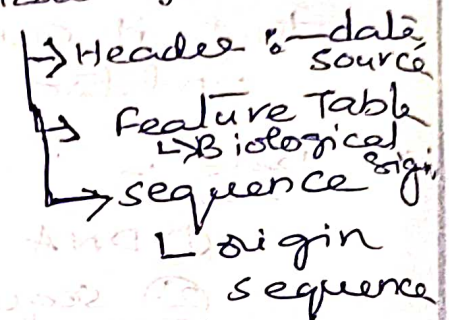
- NCBI, MD, USA - data managed.

Gene Bank Records

eg:- Flat format

→ DDBJ is located at

- National Institute of Genetics (NIG) in Japan



- DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia.

EMBL

- A comprehensive database of DNA and RNA sequence.

- established 1980

- It is maintained by EBI [European Bioinformatics Institute].

- development, maintenance of comprehensive data base of nucleotide sequence data base
- cataloging, ~~feature~~ Bibliographical, ~~texting~~ feature - location, coding Taxonomic description of source organism.

region, Biological significant site.

swissprot

- This is a curated protein sequence database that offers a high level of integration with other databases and also has very low level of redundancy.

→ It strives to provide protein sequence with a high level of annotation (for instance, the description of protein function, domain structure and post translation modification etc).

TrEMBL - Translation (Tr)

→ TrEMBL is a computer-annotated supplement of swissprot that contains all translations of EMBL nucleotide sequence entries, which is not yet integrated in swissprot.

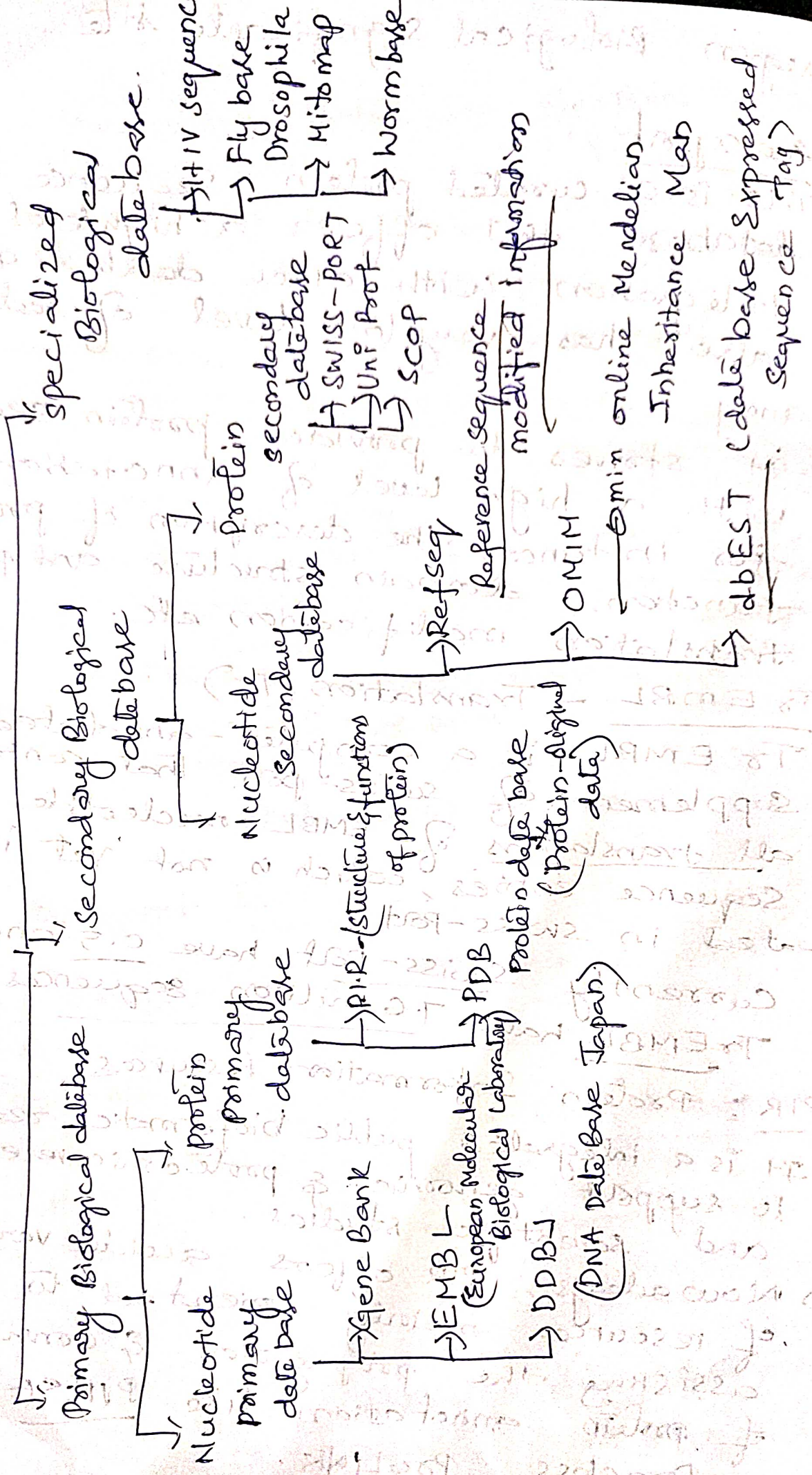
→ Currently swissprot have 0.5 and TrEMBL have 7.6 million sequences.

PIR: Protein Information Resources

→ It is an integrated public bioinformatics resource to support genomic & proteomic research and scientific studies.

→ Nowadays, PIR offers a wide variety of resources - mainly oriented to assisting the propagation & consistency of protein annotations like PIRSF, Proclass, ProLink.

Biological databases



Structural database :-

- SCOP — structural classification of protein
 - ↳ 3D structure of protein
 - ↳ Classified
- CATH — class, Architecture, Topology, Homology
 - ↳ hierarchical classification of protein
- GEO → Gene Expression Omnibus is a gene expression resource and a gene expression
- GXD → molecular abundance
- MGED → repository for gene expression data browsing query & retrieval.

Gene expression database

GXD

- gene expression data
- community resource for gene expression information

MGED

- Microarray Gene Expression data
- contain microarray data, generated by functional genomics and proteomics. Expt

metabolic → Kyoto Encyclopedia of Genes and Genome
data KEGG — pathway database — metabolic
base. LIGAND → Kyoto — chemical Research. pathway
base. LIGAND → chemical data base.

EcoCyc — E. coli. Enzyme data, repair,
Eg: — Database model organisms drug, Glycan compound

① E. coli → ① E. coli Genome Centre
(Wisconsin University USA)

② Arabidopsis thaliana → ② E. coli Index
(Birmingham. UK)
TAIR

③ Homo sapiens → HGR at NCBI, USA

④ ~~or~~ Oryza sativa → RGP (Japan)

⑤ *Drosophila melanogaster* → Fly Base

⑥ *Mus musculus* → Mouse Genome Informatics
(mouse)

⑦ Danio rerio → ZFIN (zebrafish
Information
Network
Oregon, USA.

⑧ *S. cerevisiae*. $\rightarrow$ Yeast Genome Database - Stanford
US A

Secondary data base

① PDB :- Protein Data Bank

① PDB :- Protein Data Bank

- 3D structure - macromolecules - X-ray crystallography
- Expt. data - structures - NMR
- model

→ RCSB PDB — Research Collaboration
Structural ~~B~~ data base.
tool — structure, relationship,
function, diseases causes

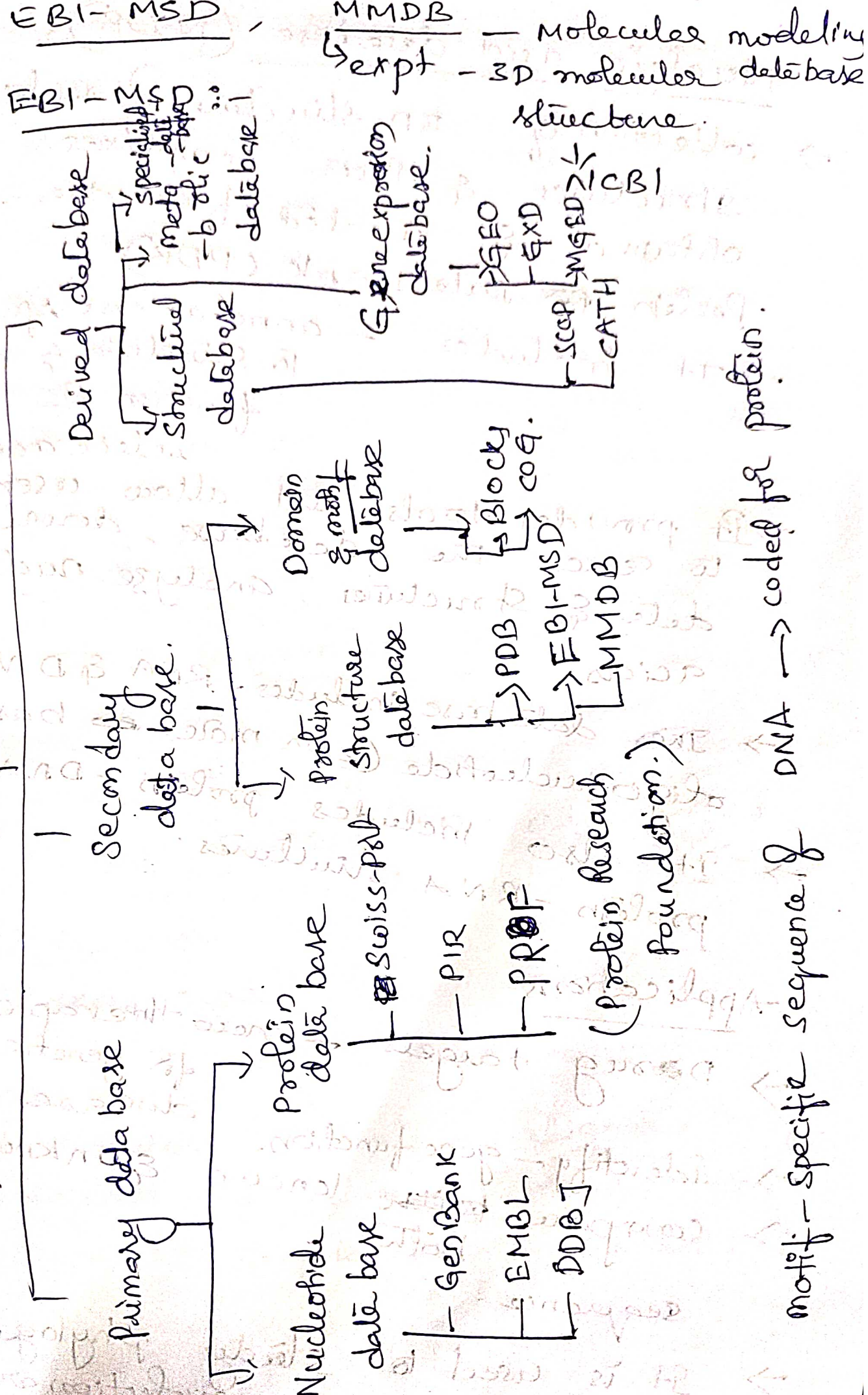
gateways to PDB files.

Swiss-Prot, NCBI, EMBL

protein data Bank.

CATH, DALI, SCOP, PDBP
databases, ~~that~~ interpret PDB files

Biological data base



② EBI-MSD

MMDB - Molecular modeling
 Expt - 3D molecular database structure.

② EBI -
 Derived database
 Structural database
 Specialized meta-database
 Gene expression database
 SCOP
 CATH
 CBI

Domain & motif database

Blocky
 COG

DNA -> coded for protein.

motif - specific sequence of

Nucleic acid database (NDB)

- collection of 3D structure of nucleic acid structures & their complexes obtained & curated from the Protein Data Bank (PDB)
- It includes - annotations specific to structure & function of nucleic acid.
- It provides tools that allow users to search the database, download data & structures, analyze nucleic acids.
- The database includes RNA & DNA oligonucleotide (2 or more ~~base~~ bases)
- It also includes protein-DNA & protein-RNA structures.

Applications

- Drug targets - new therapies for genetic diseases.
- Identify - gene function.
- Compare ~~both~~ known & unknown sequence.
- It is used to study phylogenetic analysis to analyze evolutionary

- relationships b/w organisms -
- comparing & examining their DNA & RNA sequence.
- It used to identify gene expression.
-

file format - Genebank

① Headers

- locus
- Definition
- Accession
- Version
- Source
- Keywords
- organism
- References
- Authors
- Title
- Journal
- Medline ID : all published sources
- Comment

② Features

③ sequences .

BLOCK :

- It contains multiple alignments of conserved regions in protein families.
- eg:- Google cloud Persistent Disk,

COG - clusters of Orthologous Genes

- 1997

- Identify protein sequence

→ phylogentic classification of the proteins encoded in 21 complete genomes of bacteria, archaee & eukaryotes.

→ 1,187 bacteria and 122 archaea mapped

→ uses :- Comparative evolutionary studies and enables the most specific inferences of tentative gene functions.

→ Uniport :- universal protein database

→ Ref Seq :- Reference ^{sequence} modified information

→ mito map :- Mitochondrial mapping