Sequence Databases

Jereesh A S

What is a Database/Resource?

NAR Database Issue (www.nar.oupjournals.org)

- **Collection of data in the related format**
  - structured
  - searchable (index) -> table of contents
  - updated periodically (release) -> new edition
  - cross-referenced (hyperlinks) -> links with other db
- Includes also associated tools (software) necessary for db access, db updating, db information insertion, db information deletion….
- Type and Content of Data
  - Sequence or Structure
  - Nucleic acid or protein
  - Important Biological information such as about enzyme and their metabolic pathways, mutations, diseases, drugs, images etc.

Nucleotide Databases

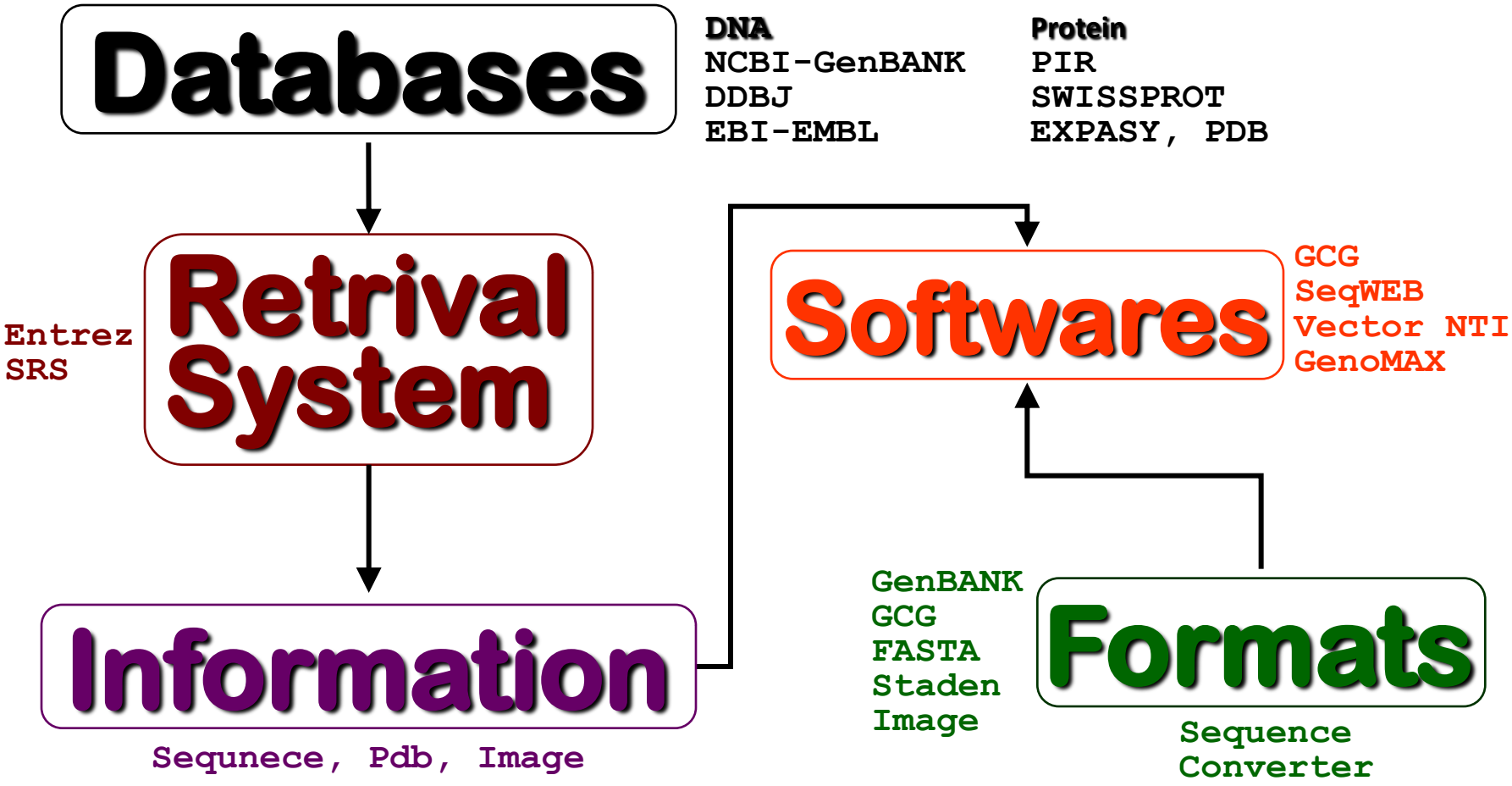
- **EMBL:** Nucleotide sequence database
- **Ensembl:** Automatics annotation of eukaryotic genomes
- **Genome Server:** Overview of completed genomes at EBI
- **Genome-MOT:** Genome monitoring table
- **EMBL-Align:** Multiple sequence alignment database
- **Parasites:** Parasite Genome databases
- **Mutations:** Sequence variation database project
- **IMGT:** Immunogenetics database, comprising-
  IMGT/LIGM- database of immunoglobulins and
  T-cell receptors, IMGT/HLA database of the human
  MHC complex and IMGT/MHC covering MHC
  complex of non-human species.

Reference site : www.ebi.ac.uk/Databases/nucleotide.html

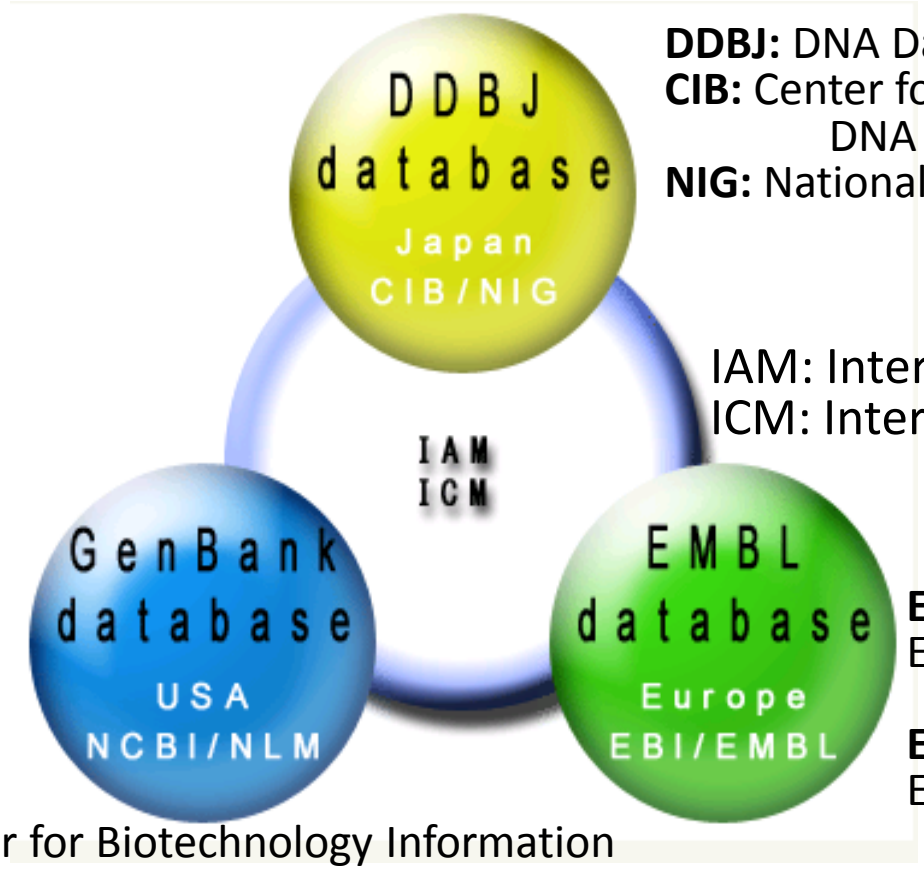
EMBL/GenBank/DDJB

- These 3 db contain mainly the same information (few differences in the format and syntax)
- Serve as **archives** containing all sequences (single genes, nucleotide, complete genomes, etc.) derived from:
  - Genome projects and sequencing centers
  - Individual scientists
  - Patent offices (i.e. USPTO, EPO)
- Non-confidential data are exchanged daily
- Currently: 2.5 $x10^7$ sequences, over 3.2 $x10^{10}$ bp;
- Sequences from > 50,000 different species;

A Sequence Retrieving and Manipulation Network



GenBank/EMBL/DDBJ International Nucleotide Sequence Database



DDBJ: DNA Data Bank of Japan
CIB: Center for Information Biology and DNA Data Bank of Japan
NIG: National Institute of Genetics

IAM: International Advisory Meeting
ICM: International Collaborative Meeting

EMBL:
European Molecular Biology Laboratory
EBI:
European Bioinformatics Institute

NCBI:
National Center for Biotechnology Information
NLM:
National Library of Medicine

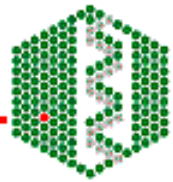
The International Nucleotide Sequence Database Collaboration

GenBank: *http://www.ncbi.nlm.nih.gov/*
National Center for Biotechnology Information (NCBI)

DDBJ: *http://www.ddbj.nig.ac.jp/*
National Institute of Genetics (NIG)

EMBL: *http://www.ebi.ac.uk*
European Bioinformatics Institute (EBI)

EMBL
European Bioinformatics Institute



ExPASy: *http://tw.expasy.org*
Expert Protein Analysis System

NCBI : GenBANK

http://www.ncbi.nlm.nih.gov

GenBank:
An annotated collection of all publicly available nucleotide and amino acid sequences.

EST database:
A collection of expressed sequence tags, or short, single-pass sequence reads from mRNA (cDNA).

GSS database:
A database of genome survey sequences, or short, single pass genomic sequences.

HTG database:
A collection of high throughput genome sequences from large-scale genome sequencing centers; including unfinished and finished sequences.

SNPs database:
A central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms.

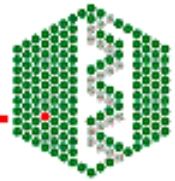
RefSeq:
A database of non-redundant reference sequences standards, including genomic DNA contigs, mRNAs and proteins for known genes. Multiple collaborations, both within NCBI and with external groups, support our data-gathering efforts.

STS database:
A database of sequence tagged sites; or short sequences that are operationally unique in the genome.

UniSTS:
A unified, non-redundant view of sequence tagged sites (STSs).

UniGene:
A collection of ESTs and full-length mRNA sequences organized into clusters, each representing a unique known or putative human gene annotated with mapping and expression information and cross-references to other sources.

EMBL
European Bioinformatics Institute



EBI:EMBL *http://www.ebi.ac.uk/services/index.html*

Nucleotide Sequence Databases

EMBL Information
EMBL Nucleotide Sequence Database information.
EMBL-Align database
EMBL-Align multiple sequence alignment database
Ensembl
Automatic annotation of eukaryotic genomes
dbEST and dbSTS Queries
Query dbEST and dbSTS.
EMEST
EMEST is a database of EST sequences.
EuroGeneIndexes
A database of EST alignments and clusters
MitBase Server
Mitochondrial DNA database server
IMGT
ImMunoGeneTics database.
EDGP
European Drosophila Genome Project server.
Parasites
Parasite Genome Databases
Mutations
Sequence variation database project.
Genomes Server
An overview of Completed Genomes at the EBI
Genome MOT
Genome Monitoring Table.

Protein Sequence Databases

SWISS-PROT
TrEMBL
InterPro

Sequence Structure Classification Databases

DSSP
Database of Secondary Structure Assignments.
HSSP
Homology Derived Secondary Structure Assignments.
FSSP
Fold Classification based on Structure-Structure Assignments.
DALI
Protein Structure Domain Dictionary
3Dee
Database of protein domain definitions.

Macromolecular Structure Databases

EBI-MSD
The EBI-Macromolecular Structure Database.

Sequence Mapping Databases

RHdb Server
Radiation Hybrid Database server.
GenomeMaps 98
Human Genome Maps 98.

DDBJ http://www.ddbj.nig.ac.jp

DDBJ (DNA Data Bank of Japan) began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG) with the endorsement of the Ministry of Education, Science, Sport and Culture. From the beginning, DDBJ has been functioning as one of the International DNA Databases, including EBI (European Bioinformatics Institute; responsible for the EMBL database) in Europe and NCBI (National Center for Biotechnology Information; responsible for GenBank database) in the USA as the two other members. Consequently, we have been collaborating with the two data banks through exchanging data and information on Internet and by regularly holding two meetings, the International DNA Data Banks Advisory Meeting and the International DNA Data Banks Collaborative Meeting.

DDBJ	15016100	22/1/02	DDBJNEW	1490104	14/3/02
DAD	945852	28/1/02	DADNEW	97212	14/3/02
SWISSPROT	105586	2/3/02	PIR	262528	11/12/01
PROSITE	1517	14/3/02	PROSITEDOC	1122	14/3/02
BLOCKS	4034	6/3/01	PRINTS	1050	6/3/01
PFAMA	2008	6/3/01	PFAMB	39228	6/3/01
SWISSPFAM	223208	6/3/01	PFAMHMM	2008	6/3/01
PFAMSEED	2008	6/3/01	PRODOM	149606	6/3/01
ENZYME	3869	29/10/01	PDB	17568	14/3/02
HSSP	15508	12/2/02	FSSP	2860	5/11/01
PATHWAY	7473	14/3/02	LENZYME	3829	13/3/02
LCOMPOUND	10158	13/3/02	SRSFAQ	10	6/3/01

Protein Databases

Protein Information Resources (PIR)

http://pir.georgetown.edu/

In 1988, The Protein Information Resource (PIR), established a cooperative effort with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID) , produces the PIR-International . Protein Sequence Database (PIR-PSD) -- a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain. The PIR-PSD, PIR-NREF, iProClass and other PIR auxiliary databases provide an integration of sequences, functional, and structural information to support genomics and proteomics research
The PIR-PSD, Current Release 71.04, March 01, 2002, Contains 283153 Entries

SWISSPROT

http://www.ebi.ac.uk/swissprot/

The SWISS-PROT Protein Knowledgebase is an annotated protein sequence database established in 1986. It is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

Protein Databases

ExPASY Molecular Biology Server

http://tw.expasy.org

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE

Protein Data Bank

http://www.rcsb.org

PROTEIN DATA BANK PDB

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

Entrez
search and retrieval system

NCBI

http://www.ncbi.nlm.nih.gov/Entrez/

Entrez *is the text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others.*

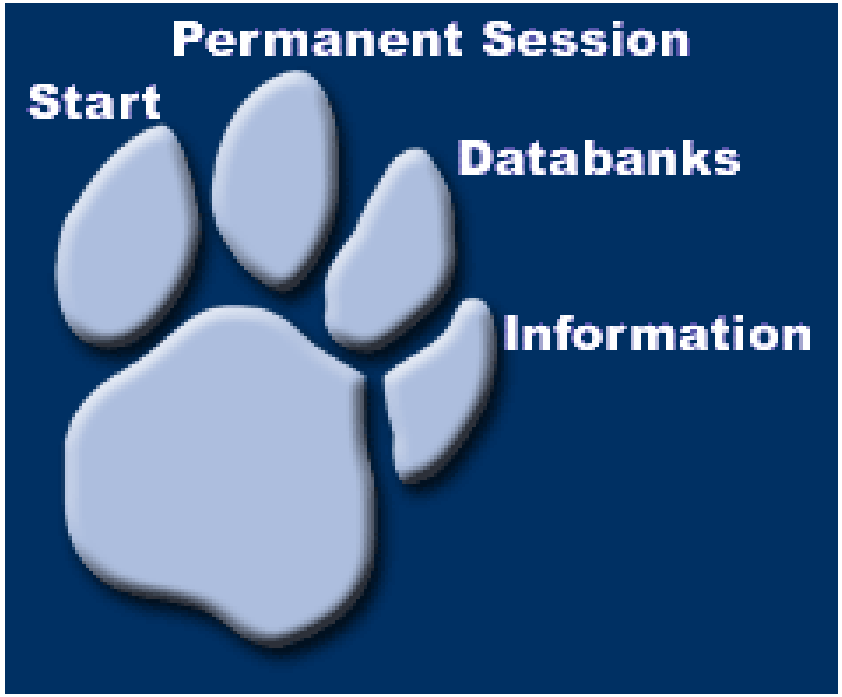
PubMed: The biomedical literature (PubMed)
Nucleotide sequence database (Genbank)
Protein sequence database
Structure: three-dimensional macromolecular structures
Genome: complete genome assemblies
PopSet: population study data sets
OMIM: Online Mendelian Inheritance in Man
Taxonomy: organisms in GenBank
Books: online books
ProbeSet: gene expression and microarray datasets
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
SNP: single nucleotide polymorphisms
CDD: conserved domains

Database Interlinking

SRS

http://srs.ebi.ac.uk/

http://srs.ddbj.nig.ac.jp/ DDBJ



http://www.lionbioscience.com/

EMBL Nucleotide Database – Europe's primary collection of nucleotide sequences is maintained in collaboration with Genbank (USA) and DDBJ (Japan)

SWISS-PROT – A complete annotated protein sequence database

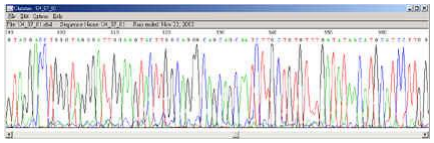
Macromolecular Structure Database - European Project for the management and distribution of data on macromolecular structures

ArrayExpress - for gene expression data

ENSEMBL - Metazoic genomes and the best possible automatic annotation.

Softwares & Sequence Formats

Program	Formats		Multiple sequence
	Default	Accept	
WWW	text file	paste & Copy	
SeqWEB	text file	paste & copy	
GCG	GCG file	FASTA GenBANK EMBL Staden SwissProt	Multiple sequence file (msf) Rich sequence file (rsf) List files (lst)
VectorNTI	*.gb *.gp	FASTA GenBANK SwissProt	FASTA GenBank SwissProt



GeneStudio
Sequence analysis software