

CONTENTS:-

- * Introduction to biological Databases.
- * What is Database.
- * Biological Database.
- * Types of Biological Databases.
- * Application of biological Databases.

INTRODUCTION TO BIOLOGICAL DATABASES

→ Emphasis is on retrieving data from the main biological databases such as GenBank.

Since the phage ϕ X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases.

WHAT IS DATABASES.

A database is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria; Databases are composed of computer hardware and software for data management.

DATA.

Data is raw, unorganized facts need to be processed.
Example: each student's test score is one piece of data.

Information:

is when data is processed, organized structured or presented in a given context so as to make it useful it's called Information.

BIOLOGICAL DATABASES

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analysis. It consists of basic units called records or entries.

A biological database can be accessed, added, retrieved, manipulated, and modified, stored, managed, connected and distributed data.

TYPES OF BIOLOGICAL DATABASES

- ≠ primary databases.
- * Secondary databases.

Primary databases: They contain biomolecular data in its original form i.e. original data in the form of primary sequence data or structural data submitted by the scientific community. The data are essentially archival in nature.

Once a database accession no. is given to an kind of data in primary databases then that data can not be changed further. Example.

1. GenBank	1. SWISS-PROT
2. EMBL	2. Pfam
3. DDBJ	3. PIR
for RNA & DNA sequences.	for protein sequences.

NUCLEOTIDE (DNA/RNA) sequences databases:

Composed of a group of nucleotide sequence entries. Data repositories that accept nucleic acid sequence data and make it freely available to the public.

GenBank, EMBL, DDBJ are principal nucleotide data bases. All the three are members of the International nucleotide sequence database consortium (INSDC) and interchange data.

DDBJ

The DNA, Data Bank of Japan is a biological data bases that collects DNA sequences. It is located at the National Institute of Genetics in the Shizuoka prefecture of Japan. It's also a member of the International nucleotide sequence Database collaboration or INSDC.

WWW.ddbj.nig.ac.jp.

Protein Sequences databases.

An array of amino acid sequence entries arranged according to the identification number. Well known protein sequence databases available are

1. Swiss-prot

4. TREMBL

2. PIR

5. UNIPROT

3. PFM

1- SWISS-PROT:- is a curated protein sequence database which strives to provide a high level of annotation (such as description of the function of proteins, its domains structure, post translational modifications variant etc) it provides all known relevant information about a particular protein.

SWISS-PROT was created in 1986 by Amos Bairoch with SWISS Institute of bioinformatics and European bioinformatics Institute (EBI).

WWW.ebi.ac.uk./swissprot/access.html.

2. PIR

The protein information resources is an integrated public resources of protein informatics that support

genomic and proteomic research and scientific discovery, it maintains the protein sequence database (PSS) an annotated protein in data base containing over 283000 sequences covering the entire taxonomic range. It assists researcher in the function identification and interpretation of protein sequence information. : <http://pir.georgetown.edu/pirwww/pirhome3.shtml>.

PROTEIN STRUCTURE Database.

Many proteins which exhibit a common evolutionary origin, show structural similarity; dissimilar proteins exhibit changes in primary, secondary, tertiary and quaternary structures, similar or dissimilar protein structure can be predicted with structure database. These databases store a collection of three dimensional structures of proteins.

THE PROTEIN STRUCTURE DATABASES

Includes, * **PDB** - (Protein Data Bank) comprises of

i- **PDBJ** - (PDB of Japan)

ii- **PDBE** (PDB of Europe)

* ~~SCOP~~ **SCOP** - (Structural Classification of Protein)

* **CATH** - (Class Architecture Topology Homology)

They all contains information generated from

i- X- crystallography.

ii- NMR- Spectroscopy Experiments.

PDB: PROTEIN DATA BANK

Understanding the shape of a molecule helps to understand how it works, it's the main primary database used for the prediction and repository of 3D structure of proteins. The data obtain from from X-ray crystallography and NMR- spectroscopy are submitted to the PDB. Then the structures are annotated as per the depositors specifications; <http://www.pdb.org/>.

SECONDARY DATABASE.

It contains data derived from results of analysed primary data i.e result of analysis of primary databases (Information that has been processed and derived from in primary database). The data in secondary database is either manually created or generated automatically. It contains some valuable information such as about mutations or Evolutionary relationship.

PROSITE

* Prosite is a protein database it consist of entries describing the protein families, domains and functional sites as well as amino acid patterns and profiles in them. They're manually curated by a team of the Swiss Institute of bioinformatics and tightly integrated into Swiss prof protein annotation.

PFAM

* Pfam is a database of protein families that includes their annotations and multiple alignments generated using hidden markov models. The generation purpose of the pfam databases is to provide a complete and accurate classification of protein families and domains.

OTHER SPECIALIZED DATABASES.

- i- HGMDB - Human gene mutation database.
- ii- OMIM - Online mutation inheritance in man (disease db)
- iii- TRANSFAC - Transcription factors database (Eukaryotic TF db)
- iv- RFAM - RNA family Database

APPLICATIONS

* the major application of these database is that they can be sharable world wide.
* they contains information about genome and

and protein of every species,
* the primary data base provide naturally from
the data so work as initial or beginning
point in any research.
* they also contains information of mutations and
evolutionary relationships.