Types of Biological Databases

Comparative Genomics Analyses Rely on Various Information in Public Databases

- General purpose sequence databases
  - Nucleotide sequence databases
  - Protein sequence databases
- Protein sequence motif and domain databases
- Protein structure databases
- Specialized genome databases
- Organism-specific databases
General Purpose Sequence Databases

Nucleotide Sequence Databases

- International Sequence Database Collaboration

1. GenBank at NCBI

2. European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database at European Bioinformatics Institute (EBI)

3. DNA database of Japan (DDBI)

Features

- All published nucleotide sequences are requested to be deposited in the one of these three databases;
- Data are exchanged among these three databases on daily basis;
- Genome sequences and annotation files are stored in the Genomes Database at NCBI, can be downloaded by anonymous ftp at ftp://ftp.ncbi.nih.gov.
General Purpose Sequence Databases

Protein Sequence Databases

Each of the three international database strives to annotate the open reading frames in each deposited sequences using computational methods.

Three issues for protein sequence databases:

1. The accuracy of prediction of protein-coding regions

2. Propagation of incorrect annotations

3. Protein classification systems
Protein Sequence Databases

The NCBI Entrez Database

- Stores translated protein sequences of nucleotide sequences from the GenBank/EMBL/DDBJ;
- Also incorporates protein sequences from SWISS-PROT and PIR;
- Each protein has unique gene identification (gi) number;
- Contain redundant sequences; thus multiple gi’s are associated with the same sequence;
- The most complete protein sequence database;
- Annotated by submitters;
- Items are linked to other NCBI databases (PubMed, taxonomy);
- NR (non-redundant) is based on Entrez contains a unique set of these sequences.
- RefSeq contains non-redundant sequences from major organisms for which sufficient data is available, in particular, sequenced genome.
Protein Sequence Databases

**PIR-PSD (Protein Information Resource-Protein Sequence Database)**

- The world's first database of classified and functionally annotated protein sequences that grew out of the Atlas of Protein Sequence and Structure (1965-1978) edited by Margaret Dayhoff;
- Produced and distributed by the Protein Information Resource in collaboration with MIPS (Munich Information Center for Protein Sequences) and JIPID (Japan International Protein Information Database);
- PIR-PSD has been the most comprehensive and expert-curated protein sequence database in the public domain for over 20 years.
- In 2002, PIR joined EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics) to form the UniProt consortium.
- PIR-PSD sequences and annotations have been integrated into UniProt Knowledgebase.
Protein Sequence Databases

**SWISS-PROT and TrEMBL Databases**

- **Swiss-Prot**: a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases;

- **TrEMBL**: a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

- Since 2002, it becomes a part of Uniprot databases.
Protein Sequence Databases

- Uniprot databases (Universal Protein Resource)
  --- Created by unification of information in three well-known protein databases through a funding from NIH.

1. PIR (Protein Information Resources);
   --- heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure

- SWISS-PROT;
  --- Created by Amos Bairock at Sweden Bioinformatics Institute

3. TrEMBL databases;
   --- Most comprehensive catalogue of information on proteins;
UniProt is comprised of three components, each is optimized for different purposes:

1. The **UniProt Knowledgebase** (UniProtKB)
2. The **UniProt Reference Clusters** (UniRef) databases
3. The **UniProt Archive** (UniParc)

Protein Sequence Databases

UniProt Knowledgebase (UniProtKB)

- In addition to capturing the core data mandatory for each UniProt entry, other available information is also added, including biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

- The UniProtKB consists of two sections:
  - UniProtKB/Swiss-Prot: manually-annotated records supported by literature and curator-evaluated computational analysis
  - UniProtKB/TrEMBL: computationally analyzed records that await full manual annotation
Protein Sequence Databases

UniProt Reference Clusters (UniRef) Databases

- The UniRef databases provide clustered sets of sequences from UniProt knowledgebase and selected UniParc records, in order to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences from view;

- The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged UniProt entries, and links to the corresponding UniProt and UniParc records.

- UniRef90 and UniRef50 are built by clustering UniRef100 sequences with 11 or more residues using an algorithm, so that each cluster is composed of sequences that have at least 90% or 50% sequence identity, respectively, to the representative sequence.
Protein Sequence Databases

UniProt Archive (UniParc)

- UniParc is a comprehensive non-redundant protein sequence collection;
- Protein sequences are loaded daily from many different publicly accessible sources to include all known sequence whose host is known;
- While a protein sequence may exist in multiple databases, and even more than once in a given database (with different identifiers), UniParc stores each unique sequence only once and assigns it a unique UniParc identifier;
- Cross-references back to the source databases are provided, and include source accession numbers, sequence versions, and status (active or obsolete).
Protein structure databases

Protein Data Bank

- Stores 3-D structures of biological macromolecules determined by experimentally (X-ray crystallography and NMR)
Protein Structure Databases

The SCOP (Structural Classification of Protein) Database

- Developed by Alexey Murzin and Cyrus Chothia.
- Hierarchical classification of structural domains of individual PDB entries;
- The SCOP is organized as a tree structure with the hierarchy of

- **Class**: According to the arrangement of secondary structure

  - **Fold**: Proteins have same major secondary structures in same arrangement with the same topological connections

  - **Superfamily**: Low sequence identities but structures suggest that a common evolutionary origin is probable.

  - **Family**: Cluster of homologous proteins
SCOP Classes

1. **All alpha proteins** [46456] (226) [protein domains] (folds)
2. **All beta proteins** [48724] (149)
3. **Alpha and beta proteins (a/b)** [51349] (134)
   *Mainly parallel beta sheets (beta-alpha-beta units)*
4. **Alpha plus beta proteins (a+b)** [53931] (286)
   *Mainly antiparallel beta sheets (segregated alpha and beta regions)*
5. **Multi-domain proteins (alpha and beta)** [56572] (48)
   *Folds consisting of two or more domains belonging to different classes*
6. **Membrane and cell surface proteins and peptides** [56835] (49)
   *Does not include proteins in the immune system*
7. **Small proteins** [56992] (79)
   *Usually dominated by metal ligand, heme, and/or disulfide bridges*
8. **Coiled coil proteins** [57942] (7)
   *Not a true class*
9. **Low resolution protein structures** [58117] (24)
   *Not a true class*
10. **Peptides** [58231] (116)
    *Peptides and fragments. Not a true class*
11. **Designed proteins** [58788] (42)
    *Experimental structures of proteins with essentially non-natural sequences.*
Protein Classification Databases

The CATH Database

- Developed by Janet Thornton and Christine A. Orengo.
- Hierarchical classification of structural domains of individual PDB entries;
- The CATH is organized as a tree structure with the hierarchy of:
  
  **Class:** According to the arrangement of secondary structure

  **Architecture:** orientation of secondary structure elements

  **Topology:** topological connections between secondary structure elements

  **Homologous superfamily:** Cluster of homologous proteins
Protein sequence motif: a set of conserved amino acid residues that are important for protein function and are located in a short distance from one another.

The EF-hand motif

| -helix | ----- | ------- | loop1 | ------- | --helixII |-

\[ \text{En}^{\text{**nn**}}-nX^{\text{**Y}}-Z^G-I^x^{\text{**zn**}}n{\text{**nn**}}n \]

Legend: E = glutamate; n = hydrophobic residue; * = any residue; X = first calcium ligand; Y = second calcium ligand; Z = third calcium ligand; G = glycine; # = fourth calcium ligand, provided by a backbone carbonyl; I = isoleucine (although other aliphatic residues are also found at this position); -X = fifth calcium ligand; -Z = sixth and seventh calcium ligands, provided by a bidentate glutamate or aspartate residue.
Protein Sequence Motif Databases
The PROSITE database

- Created by Amos Bairock at Sweden Bioinformatics Institute
- Consists of a large collection of biologically meaningful signatures that are described as patterns or profiles.

- Release 20.17, of 24-Jul-2007 contains
  --- 1489 documentation entries
  --- 1319 patterns
  --- 739 profiles
Protein Sequence Motif Databases

The PRINTS Database

- PRINTS is a compendium of protein fingerprints.
- A fingerprint is a group of conserved motifs used to characterize a protein family;
- Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space.
- Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs: the database thus provides a useful adjunct to PROSITE.