Primary databases of proteins

- One of the hallmarks of modern genomic research is the generation of enormous amounts of raw sequence data.
- The very first challenge in the genomics era is to store and handle the staggering volume of information through the establishment and use of computer databases.
- A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.
- A simple database might be a single file containing many records each of which includes the same set of information.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.

Importance of protein databases

Huge amounts of data for protein structures, functions and particularly sequences are being generated. Searching databases is often the first step in the study of a new protein. It has the following uses:

- Comparison between proteins or between protein families provides information about the relationship between proteins within a genome or across different species and hence offers much more information that can be obtained by studying only an isolated protein.
- Secondary databases derived from experimental databases are also widely available. These databases reorganize and annotate the data or provide predictions.
- The use of multiple databases often helps researchers understand the structure and function of a protein.

The PRIMARY databases hold the experimentally determined protein sequences inferred from the conceptual translation of the nucleotide sequences. This of course is not experimentally derived information but has arisen as a result of interpretation of the nucleotide sequence information and consequently must be treated as potentially containing misinterpreted information. There are a number of primary protein sequence databases and each requires some specific consideration.

The Universal Protein Resource (UniProt)

- The Universal Protein Resource (UniProt) provides a stable, comprehensive, freely accessible, central resource on protein sequences and functional annotation.
- The UniProt Consortium is a collaboration between the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB).
- The core activities include manual curation of protein sequences assisted by computational analysis, sequence archiving, development of a user-friendly UniProt website, and the provision of additional value-added information through cross-references to other databases.
- UniProt is comprised of four major components, each optimized for different uses: the UniProt Knowledgebase, the UniProt Reference Clusters, the UniProt Archive and the UniProt Metagenomic and Environmental Sequences database.
- UniProt is updated and distributed every three weeks, and can be accessed online for searches or download at http://www.uniprot.org.

The Uniprot Knowledgebase (UniProtKB)

UniProtKB consists of two sections: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL.

- The UniProtKB/Swiss-Prot contains manually annotated high quality records with information extracted from literature and curator-evaluated computational analysis. Sequences for which novel functional, structural and/or biochemical data have been published are assigned priority.
- The annotation in UniProtKB consists of the description of the following: function(s), enzyme-specific information, biologically relevant domains and sites, post-translational modifications, subcellular location(s), tissue specificity, developmentally specific expression, structure, interactions, splice isoform(s), diseases associated with deficiencies or abnormalities.
- UniProtKB/TrEMBL contains computationally analyzed records enriched with automatic annotation and classification. It contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases, the sequences of PDB structures and data derived from amino acid sequences that are directly submitted to the UniProt Knowledgebase or scanned from the literature.

Importance of Databases

- Databases act as a store house of information.
- Databases are used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- It allows knowledge discovery which refers to the identification of connections between pieces of information that were not known when the information was first entered. This facilitates the discovery of new biological insights from raw data.
- Secondary databases have become the molecular biologist's reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.
- It helps to solve cases where many users want to access the same entries of data.
- Allows the indexing of data.
- It helps to remove redundancy of data.