DEPARTMENT OF COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

PROTEIN DATABASES: DRAFT

Objective: To familiarize protein databases there by giving a learning experience to retrieve information from databases such as Uniprot and PDB a sequence database a structure database respectively.

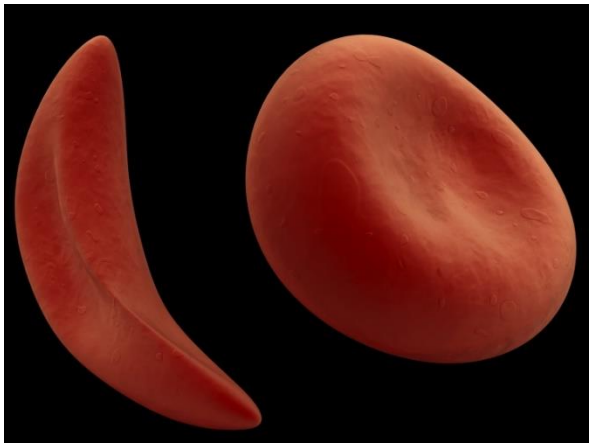
UNIPROT

Universal Protein Knowledgebase (UniProt) consortium is a single centralized resource for providing sequences and functional information of a protein. The central database will have two sections, corresponding to the familiar Swiss-Prot (fully manually curated entries) and TrEMBL (enriched with automated classification, annotation and extensive cross-references). The primary mission of the consortium is to support biological research by maintaining a high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community.

The UniProt databases consist of three database layers :(i) The UniProt Archive (UniParc) provides a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data. (ii) The UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation. (iii) The UniProt NREF databases (UniRef) provide non-redundant data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions.

Exercises

Sickle cell anaemia is a genetic disorder due to single point mutation in the β-globin chain of Haemoglobin. Haemoglobin possess higher structural conformation consists of four subunits two identical alpha chains and two identical beta chains. The protein function as a carrier protein to transport oxygen to various parts of the system. In sickle shaped



a) Sickle shaped and normal RBC

b) Normal RBC (Discoid shape)

RBC the hydrophilic amino acid glutamic acid to be replaced with the hydrophobic amino acid valine at the sixth position of the β-globin chain. To retrieve the sequence information of beta globin chain of Haemoglobin protein from Uniprot

I) Retrieve the information of beta globin chain of Haemoglobin using the Uniprot ID P69892

1. Write down the contents of feature table.
2. What are the molecular and biological function of the protein? Cite with one reference.
3. What are the amino acid modification of these proteins?
4. How many binary interactions of these proteins are found?
5. How many protein structures are found and their PDB ID?
6. How many isoforms of protein are there?
7. What is the length of the amino acid composition?
8. What are the structural differences of these proteins with references?
9. What is the chromosome location of Beta globulin chain?
10. Find the experimental annotation of the protein?

Protein Data Bank

The Protein Data Bank (PDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. Understanding the shape of a molecule deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

Retrieve the information using the database PDB by searching the keyword Human Haemoglobin

1. How many structures are deposited in the PDB for Human Haemoglobin?

2. Write down any 5 PDB ID of the human haemoglobin and cite their references?

3. How many citations are there in the PDB associated with human Haemoglobin protein structure?

4. Write down the structural properties of the haemoglobin with PDB ID: 1HAB

(a) Experimental methods

(b) Date of submission

(C) Resolution of the protein structure

(d) PubMed ID

(E) Total structure weight

(D) Protein chains

(F) Sequence length

(G)Organism

(H) Ligands bounded to proteins

(I) References