**Objective: Explore different types of biological databases and classify them.**

**Task: Choose any three databases (Primary, Secondary and Specialized databases)**

**Identify and describe their data type, format and purpose.**

Bioinformatics relies heavily on databases to store, analyze and interpret biological data. The different types of bioinformatics databases are:

**PRIMARY DATABASES** – They store raw, unprocessed data directly from experimental methods. For example:

|  |  |  |  |
| --- | --- | --- | --- |
| **PRIMARY DATABASES** | **DATATYPE** | **FORMAT** | **PURPOSE** |
| **GenBank** | Nucleotide sequences, Annotation, bibliographic and biological information. | Flat file format | Provide and encourage access within scientific community to the most up to date and comprehensive DNA sequence information. |
| **EMBL( European Molecular Biology Laboratory)** | Nucleotide sequences, protein sequences and related information. | Flat file format | To perform fundamental research, offer vital service to scientists, train scientist and students and engage in technology transfer and industry relations. |
| **DDBJ( DNA Databank of Japan)** | Nucleotide sequence data including WGS (Whole Genome Shotgun), RNA, cDNA, synthetic sequences, environmental sequence data and data from Next Generating Sequencing platforms. | Flat file format | Serves as repository for nucleotide and Amino Acid sequence data and also a key player in INSDC (International Nucleotide Sequence Database Collection). |
| **PDB( Protein Databank)** | Stores and shares 3D structure of biological macromolecule including (Protein and Nucleic Acid) along with related information. | Text file format | Serves as central resource for structural information, enabling researchers to study protein structure, understanding their function and develop new technologies. |

**SECONDARY DATABASES**- They are derived from primary databases which focuses on processed or interpreted data such as annotations, motifs or protein families. For example:

|  |  |  |  |
| --- | --- | --- | --- |
| **SECONDARY DATABASES** | **DATATYPE** | **FORMAT** | **PURPOSE** |
| **UniProt** | Freely accessible database of protein sequences and functional information. | Flat text, FASTA, XML, RDF and GFF. | It is used for storing and interconnecting information from large and disparate sources. |
| **Pfam** | It is a curated database of protein families, domains, repeats, motifs, coiled-coil and disorders. | It is an ASCII file. | It aids in genome annotation, protein structure and functional analysis, and evolutionary studies. |
| **PROSITE** | It is a database of documentation entries that describe protein domains, families, and functional sites derived from multiple sequence alignment. | Three ASCII text files, including a DAT file for motif matches. | It serves to identify protein families, domains, and functional sites by using patterns and profiles. |
| **PRINTS (Protein Family Fingerprints)** | It is a collection of protein fingerprints, which may be used to assign family and functional attributes to uncharacterized sequence. | PDF, JPEG, TIFF and EPS. | It serves as diagnostic resource for identifying and characterizing newly determined protein sequences. |
| **SWISS-PROT** | It is a curated protein sequence database with a high level of annotations such as functions, domain structure, post-translational modifications and variants. | Flat file format. | To provide a curated, high quality and non-redundant protein sequence database with detailed annotations. |

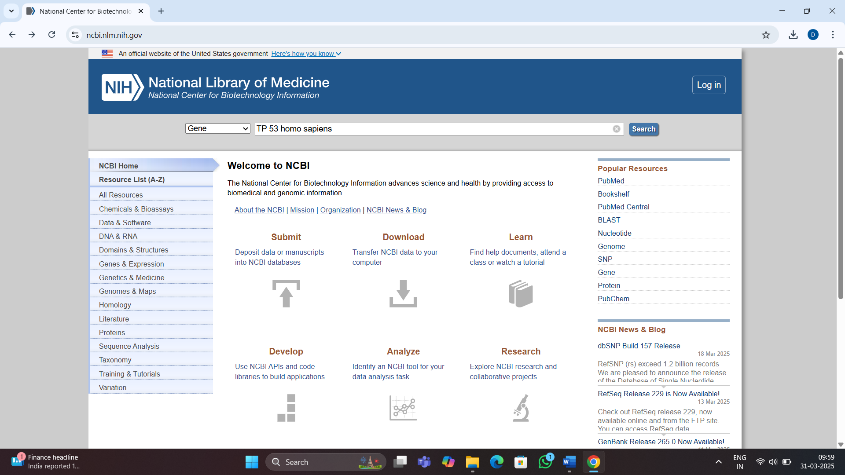
**SPECIALISED DATABASES**- These databases are designed to serve a specific research interest, concentrating on a particular organism, datatype and biological process. For example:

|  |  |  |  |
| --- | --- | --- | --- |
| **SPECIALIZED DATABASES** | **DATATYPE** | **FORMAT** | **PURPOSE** |
| **Flybase (Drosophila)** | It is a web portal for genetic and genomic information on *Drosophila melanogaster* and related species, with data types including genes, alleles, phenotypes, aberrations, sequences, stocks, images, and more | XML, ASCII flat files, FASTA, GFF and GTF files. | It has information on gene sequences, alleles, mutations, genome and amino acid sequences of *Drosophila melanogaster.* |
| **HIV sequence database** | It encompasses both nucleic acid and amino acid sequences, crucial for understanding the virus’s structure, evolution, and drug resistance and is used to classify HIV into groups and subtypes. | FASTA format. | It helps to collect, curate and analyse HIV sequences, aiding in research on drug resistance, vaccine development and understanding viral evolution. |
| **RDP(Ribosomal Database Project)** | It primarily deals with aligned and annotated ribosomal RNA sequences, derived phylogenetic trees, and related analysis tools. | GenBank-formatted files. | It provides the research community with aligned and annotated rRNA gene sequence data, analysis services, and tools, including a phylogenetically consistent taxonomic framework. |
| **KEGG (Kyoto Encyclopedia of Genes and Genomes)** | It is an ontology database containing hierarchical classifications of various entities including genes, proteins, organisms, diseases, drugs, and chemical compounds. | KGML (KEGG Markup Language). | It is used for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. |
| **REACTOME** | It is an open-source, open access, manually curated and peer-reviewed pathway database. | BioPAX, PSI-MITAB, SBML, and tab-delimited text files. | It provides tools for the visualization, interpretation and analysis of pathway knowledge to support basic and clinical research, genome analysis, modelling, systems biology and education. |

**Retrieve one dataset from each and explain how to extract relevant information**.

1.PRIMARY DATABASE(GenBank)

* Go to the NCBI website.
* Select gene from drop down menu and type your gene of interest on search bar ( TP (Tumor Protein )53 *Homo sapiens).*

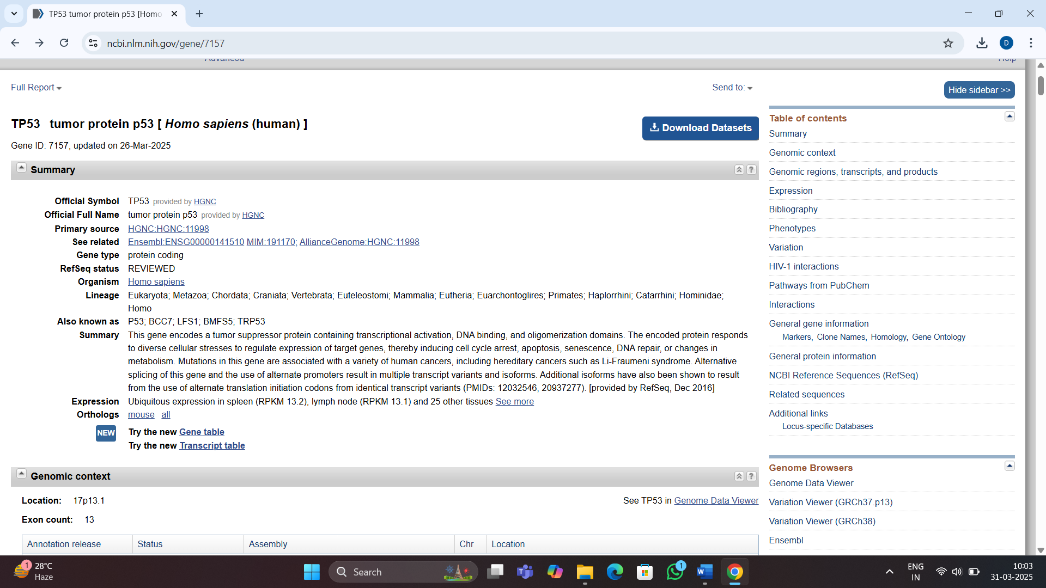


* Select the file you want to open, here we are focusing on studying the gene sequence of *Homo sapiens* .

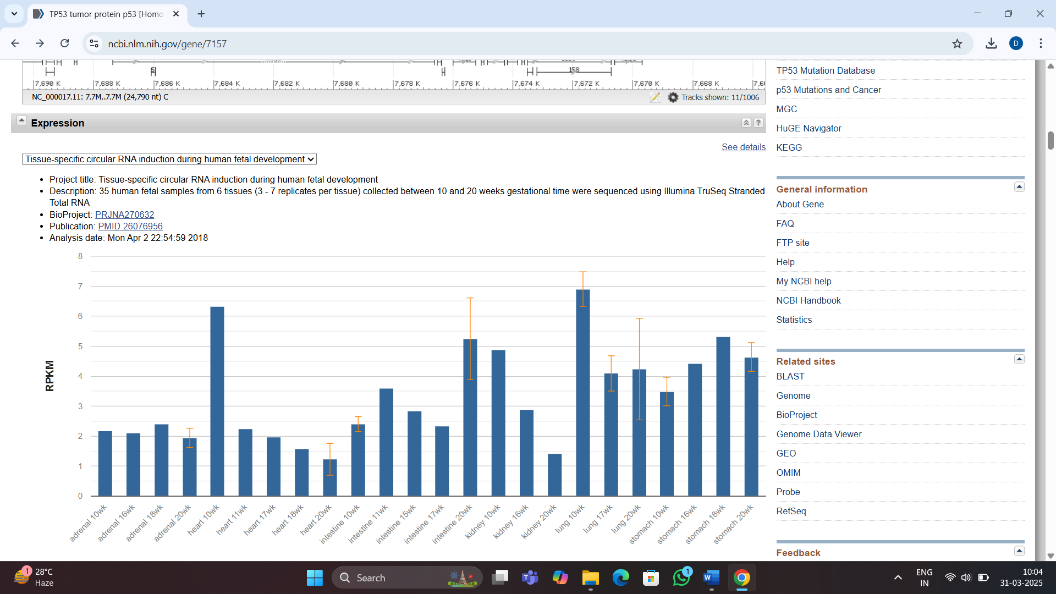
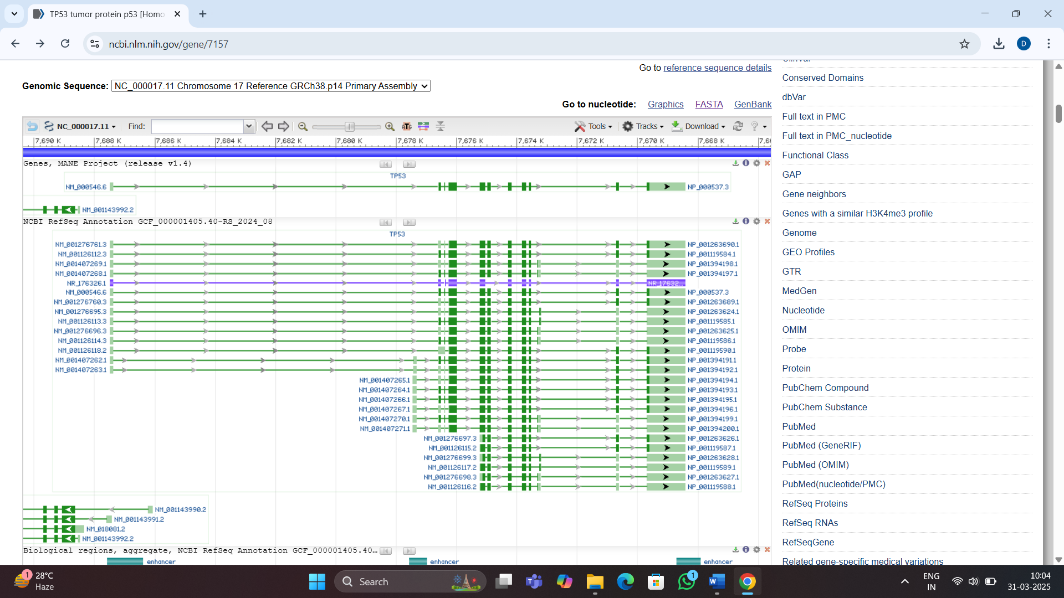
A screenshot of a computer

AI-generated content may be incorrect.

* You can then analyze the summary, genomic content, expression, bibliography etc.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* You can then explore the genomic sequence from GenBank option of Genomic Sequence to find the whole gene sequence, CDS regions, messenger RNA sequences, etc. , which helps in understanding gene function, variations, comparative study, gene discovery, develop personalized medicine and discover new treatments.

A screenshot of a computer

AI-generated content may be incorrect.

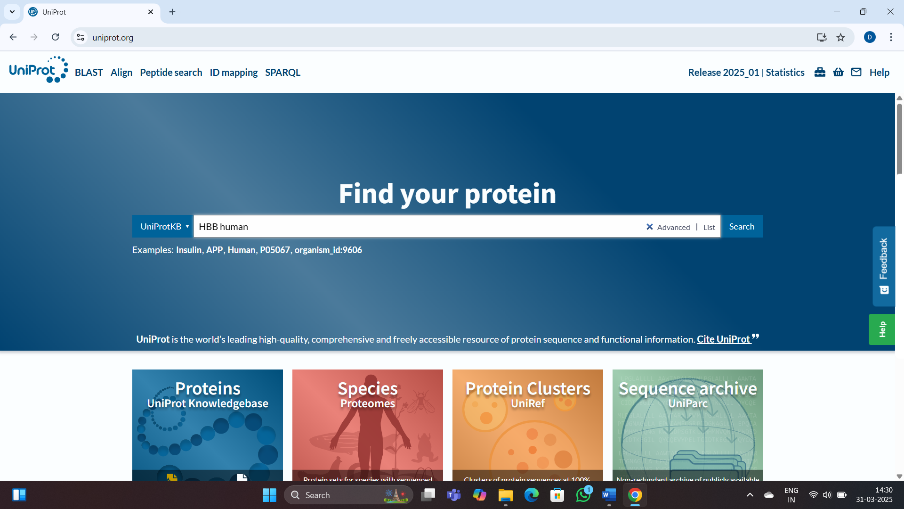
* You can then get the FASTA sequence from the GenBank for sequence analysis with improved accuracy in genomics, transcriptomics, proteomics studies.

A screenshot of a computer

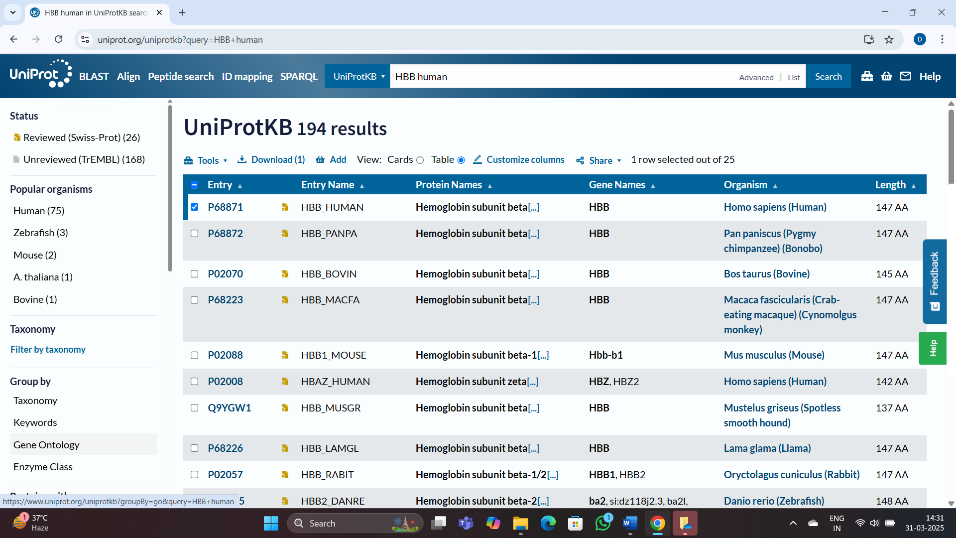
AI-generated content may be incorrect.

2. SECONDARY DATABASE (UniProt)

* Go to the UniProt website and search your protein of interest ( HBB (Haemoglobin subunit Beta) in humans).



* Select the protein sequence you want to study.



* Under function, we can study how protein plays an important role in transporting oxygen, how mutations lead to genetic disorder and how it is highly conserved in different species.

A screenshot of a computer

AI-generated content may be incorrect.

* Gene ontology provides information on the transportation of oxygen, biosynthesis of haemoglobin, differentiation of RBCs, how HBB is associated with RBC membrane and binds oxygen.

A screenshot of a computer

AI-generated content may be incorrect.

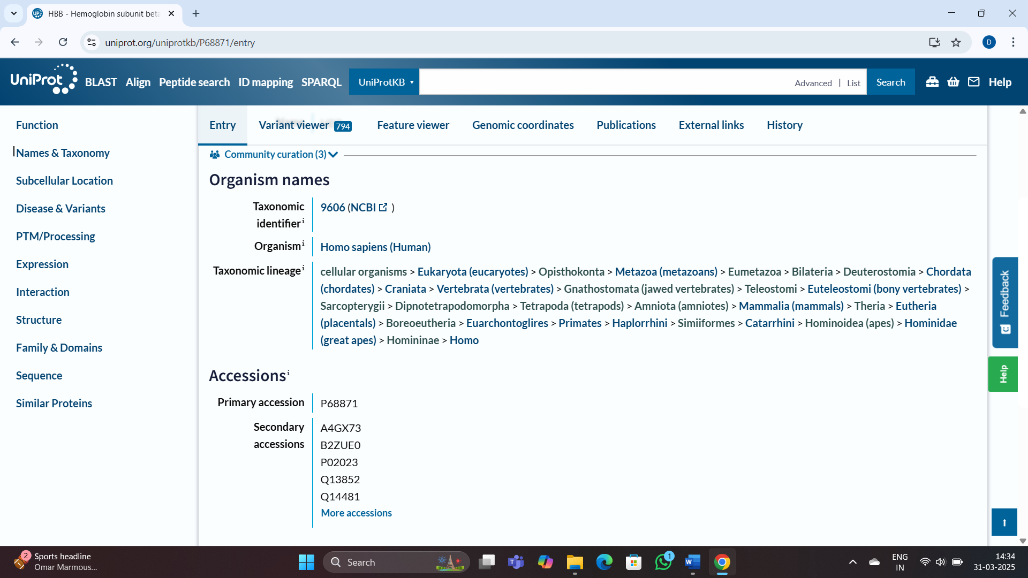
* Enzyme and pathway database tells about the globin protein and it’s oxygen binding activity, the oxygen transport pathway, the RBC differentiation pathway, interactions with other globin proteins and molecules.

A screenshot of a computer

AI-generated content may be incorrect.

* Names and taxonomy gives insights of it’s alternative names, it’s classification, accession number, organism it is obtained from and the lineage.

A computer screen shot of a computer

AI-generated content may be incorrect.

* Subcellular location specifies the places the gene is located so that we can understand the prime functions and repercussion of it’s possible mutations. The gene ontology gets further explained here, as we find the exact location to study the functions. A computer screen shot of a cell

  AI-generated content may be incorrect.
* Disease and variants, specifies about the different genetic disorders like Beta Thalassemia, Sickle cell anaemia,HEIBAN ; also about the variants of the gene which could to different forms of disorders, different mutations leading to a gennetic disorder and the mechanism of disease which helps in genetic counselling.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

* Post translational modification helps us understand the regulation of the function in normal state and the effects if the modifications are absent, the interactions could be hampered, and the protein would lose stability. It helps discover biomarkers for disease diagnosis.

A screenshot of a computer

AI-generated content may be incorrect.

* The tissue specific expression tells about the transcription factors and proteins it is associated with, and the interactions which leads to the expression of the functions.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

* Structure classifies the size, exonic and intronic regions, promoter sequences, regulatory elements, helix diameter, transcription start site to understand mutations.

A computer screen shot of a structure

AI-generated content may be incorrect.

* It is a member of globin gene family, which characterizes it’s functional and protein domains.

A computer screen with a computer screen

AI-generated content may be incorrect.

* Sequence helps classify point, frameshift, deletion or other mutations which leads to disorders.

A computer screen with a message

AI-generated content may be incorrect.

* Polymorphism tells how SNPs, insertion and deletion, copy number variations could lead to disorder. Sequence similarity helps identify the conserved sequences, homology among species and other proteins and phylogenetic studies.

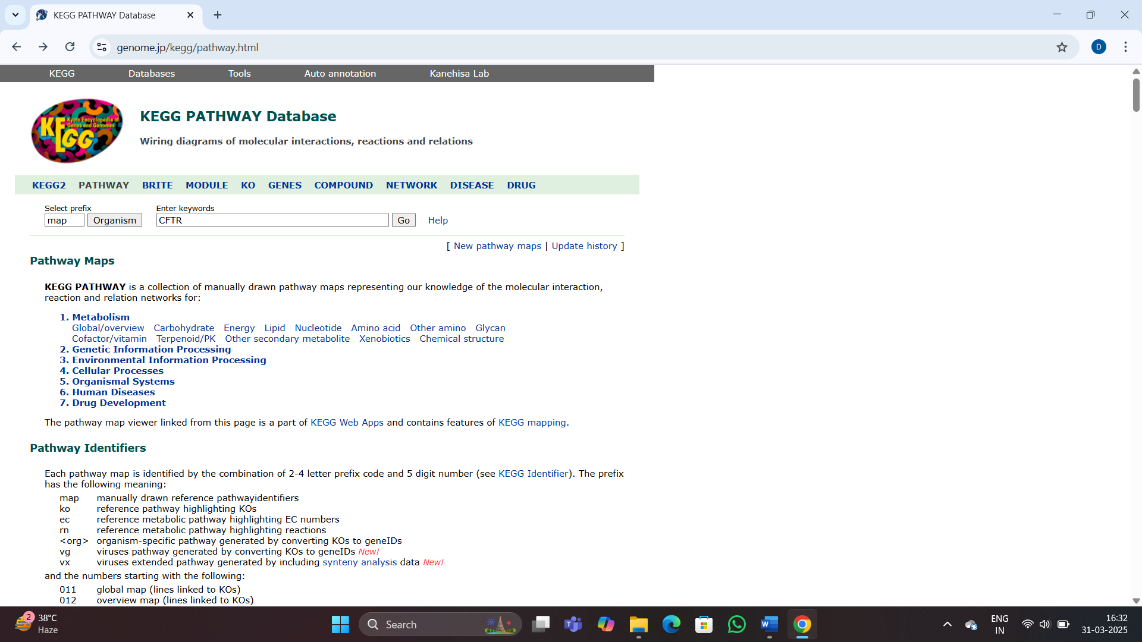
A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**3. SPECIALIZED DATABASE (KEGG)**

* Go to the KEGG database and enter gene of interest ( CFTR ; Cystic Fibrosis Transmembrane Conductance Regulator) whose pathway you want to study.



* The results shows different pathways the gene is involved in, click on the one you want to study( cyclin AMP signaling pathway).

A screenshot of a computer

AI-generated content may be incorrect.

* The description elaborates about the pathway and the pathway map shows how the gene is directly or indirectly regulating various gene expression and functions.

A screenshot of a computer

AI-generated content may be incorrect.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

* The references of various studies or experiments are mentioned which led to creation of the pathway map and in the gene ontology section, similar pathway regulated by different gene in other organisms can be studied.A screenshot of a computer

  AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**COMPARE HOW SEQUENCE DATA IS STRUCTURED IN NCBI vs UniProt**

|  |  |  |
| --- | --- | --- |
| **S.NO.** | **NCBI** | **UniProt** |
| 1. | It focuses on nucleotide sequences (DNA/RNA) and related genomic information. | It focuses on protein sequences and functional information. |
| 2. | Some databases involved are:  GenBank- Primary nucleotide sequence database.  RefSeq- A curated, non-redundant database of reference sequences.  NCBI Structure- It has experimentally resolved structures of proteins and nucleotides derived from PDB. | Some databases involved are:  UniProtKB- A curated database of protein sequences and functional annotations.  UniRef- Non-redundant sequence clusters at different levels of sequence identity.  UniParc- A comprehensive archive of all known protein sequences. |
| 3. | It uses variety of formats for storing sequence data, including FASTA and GenBank formats. | It uses an ASCII flat file format for its main distribution. |
| 4. | It maintains the Sequence Read Archive for storing and analyzing Next-Generation Sequencing data. | Entries are structured to include sequence information, functional annotations, cross-references to other databases, and experimental data. |
| 5. | It provides a comprehensive resource for nucleotide sequence data, genomic information, and related tools and services. | It provides a comprehensive and integrated resource for protein sequence and serves as a central hub for protein knowledge. |
| 6. | It assigns unique identifiers to each sequence, which can be used to retrieve and manipulate the sequence data. | It assigns unique identifiers to each protein, which can be used to retrieve and manipulate the protein data. |
| 7. | It provides information on sequence features, such as coding regions, non-coding regions, and regulatory elements. | It provides information on protein features, such as domains, motifs, and post-translational modifications. |
| 8. | It offers tools for performing sequence alignments, which can be used to compare and analyze sequences. | It offers annotations on sequences, which can include information on protein function, structure and interactions. |