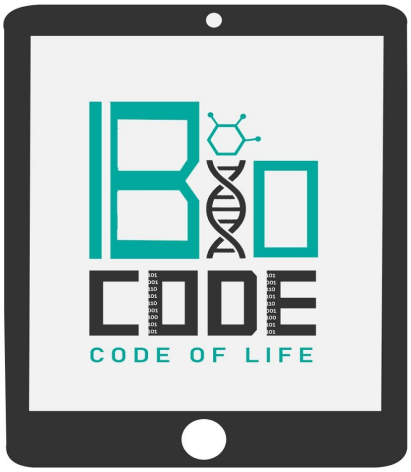

LEARN PYTHON & R FOR BIOINFORMATICS

Prerequisite terminologies:

In order to have a thorough understanding of our main topic, you should have the basic concept of the following terminologies:

1. Secondary Database.
2. Protein annotation.
3. Protein sequences.

Introduction:

UniProt is a secondary database that contains information about the protein sequences, annotations, etc. This information is derived from genome sequencing projects. It specifically contains the information about the proteins only, so that one can access, retrieve, analyze and download the publicly available data, of any protein, from UniProt in order to utilize the information in other various researches. It contains a large amount of information about the biological function of proteins derived from the research literature. UniProt was also known as “SwissProt” in the earlier days. There are 4 basic sub-databases within UniProt including UniRef, UniParc, Proteomes, and UniProtKB.

Steps:

- Click on the link below to visit the homepage of UniProt: https://www.uniprot.org/
- **UniProtKB** stands for 'UniProt knowledge base', having 2 basic subsets (or sub-databases):
    - **Swiss-Prot:**
        - ➔ It contains the information carefully curated by the researchers manually or the information from the literature, that is highly experimentally validated information.
    - **TrEMBL:**
        - ➔ It contains the information which is usually not manually curated by the researchers rather it contains the information that is automatically annotated by the TrEMBL pipelines in UniProt.
        - ➔ TrEMBL stands for Translated European Molecular Biology Laboratory.
- **UniRef** is a reference cluster of those proteins that are in a single clustering form.
    - **For example,** if you have a protein sequence having different isoforms as well, UniRef will cluster all the isoforms of the required protein in one record so you can access all the isoforms in one single record rather than search each isoform one by one, manually.
- **UniParc** is the comprehensive database that contains non-redundant records of a protein sequence.
    - **For example,** if you are working on a specific protein, on which other research has also been done by other researchers which has generated redundant data of that protein within the UniProt, so UniParc will remove the redundancy and provide you the best and validated data about that protein.
- **Proteomes** is the most important database of UniProt. It allows you to analyze the entire proteome of an organism that you're trying to analyze or any other organism that you think of.

    - UniProt basically converts the genomic data into proteomic information of an organism, which can either be done manually or using any computational method, and that information is then stored in the Proteomes database of UniProt.
- **Supporting Data** which contains information about ‘Literature and Citations’, ‘Taxonomy’, ‘Subcellular locations’, ‘Cross-ref database’, ‘Diseases’ and ‘Keywords’.

Summary:

In this introductory video of UniProt, we learned how to use the UniProt database and for which purposes we can use it. We also came to know about the different sub-databases hosted by UniProt that provide precise information about a protein.