Experiment 2: Exploring Protein Sequence Databases

AIM

To explore protein sequence databases (UniProt, PROSITE, PFAM) and retrieve comprehensive information about proteins, including sequence data, functional annotations, domain structures, and evolutionary relationships.

THEORY (For more detail, please refer to the PowerPoint slides)

Protein sequence databases serve as comprehensive repositories of biological information, containing millions of protein sequences along with their functional annotations, structural data, and evolutionary relationships. These databases are fundamental resources in bioinformatics, enabling researchers to understand protein function, predict structural features, and trace evolutionary patterns across species. The integration of experimental data with computational predictions has made these databases indispensable tools for modern biological research, drug discovery, and biotechnology applications.

UniProt (Universal Protein Resource) represents the most comprehensive protein sequence and annotation database available today, accessible at https://www.uniprot.org/. This database integrates protein sequences with extensive functional information, including enzyme classification, subcellular localization, disease associations, and literature references. UniProt combines three distinct databases: UniProtKB (the central hub), UniRef (sequence clusters), and UniParc (sequence archive), providing both manually curated high-quality annotations and automatically generated entries. The database contains detailed information about protein sequences, their functions, taxonomy, subcellular locations, post-translational modifications, and associated diseases, making it the primary reference for protein-related research.

PROSITE, available at https://prosite.expasy.org/, specializes in identifying and cataloging protein domains, families, and functional sites through pattern recognition and profile analysis. This database focuses on protein motifs, patterns, and functional signatures that are conserved across protein families, enabling the prediction of protein function based on sequence similarity. PROSITE employs regular expressions and position-specific scoring matrices to identify functional domains and active sites, making it particularly valuable for functional annotation of newly discovered proteins.

PFAM (Protein Families), accessible at http://pfam.xfam.org/, represents a comprehensive collection of protein families characterized by multiple sequence alignments and hidden Markov models. This database organizes proteins into families based on evolutionary relationships and shared structural domains, providing insights into protein evolution and function. PFAM contains protein domain families, carefully curated alignments, phylogenetic trees, and domain architecture information, serving as a crucial resource for understanding protein evolution and domain organization.

Exercise 1: UniProt Information Retrieval

Task: Retrieve comprehensive information about human "G-protein coupled receptor kinase 2 (GRK2)"

Steps:

1. Go to https://www.uniprot.org/
2. Search for "G-protein coupled receptor kinase 2 (GRK2)"
3. Click on the entry from humans.
4. Complete the data sheet below

Data to Record:

1. Accession number
2. Protein name
3. Gene name
4. Organism
5. Sequence length (in amino acids). Molecular weight (in Da)
6. Function (one sentence)
7. Subcellular location
8. Download the sequence. Save the sequence as FASTA file with a proper name.
9. Examine the Features and take a screenshot. Save the file with a proper filename.

Exercise 2: PROSITE Motif Identification (Using ScanProsite)

Task: Identify functional motifs in human GRK2 using ScanProsite

Steps:

1. Go to https://prosite.expasy.org/scanprosite/
2. Copy and paste the sequence of human G-protein coupled receptor kinase 2 (GRK2) (you downloaded it in the previous exercise)
3. Select the appropriate options and then START THE SCAN
4. Record the data.
5. Go back and in STEP 2 Select "Run the scan at high sensitivity (show weak matches for profiles)"

Data to Record:

- Number of significant hits by profiles and their details
- Number of significant hits by patterns and their details
- Take a screenshot of the graphical view and save the image with a proper filename.
- After running Step 5, what difference did you notice in the output? Record your observation.

Exercise 3: PFAM Domain Analysis

Task: Analyze the sequence of the human "G-protein coupled receptor kinase 2 (GRK2)"

Note: PFAM is now hosted by InterPro (http://pfam.xfam.org/)

Steps:

1. Go to https://www.ebi.ac.uk/interpro/search/text/
2. Search for "insulin"
3. Click on P25098 (Uniprot accession for human G-protein coupled receptor kinase 2 (GRK2)
4. Explore families and domains. Find the Pfam results in the "Representative families" column
5. Click on the Pfam results (there are three), and record the following data for each result
6. Within each Pfam result page, go to the Taxonomy tab and view key species.

Data to Record:

- Pfam accession
- Family name and description
- In the Taxonomy > Key Species, how many similar proteins are in *Arabidopsis thaliana* and the Fruit Fly?

DISCUSSION Section

Please address the following in your discussion section:

1. Compare and contrast the types of information provided by UniProt, PROSITE, and PFAM for the same protein (GRK2).

    - How do these databases complement each other in providing a comprehensive understanding of protein structure and function?
    - Discuss why it might be necessary to consult multiple databases rather than relying on a single resource when studying protein.

2. Consider a scenario where you have identified a novel protein sequence with unknown function. Based on your experience with these three databases, outline a systematic approach for functional annotation. Discuss how the information from UniProt, PROSITE, and PFAM/InterPro could be integrated.