Biological Data

Data Science

Group project - Family model

*"A **protein domain** is a conserved part of a given protein sequence and tertiary structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded."* (Wikipedia).

The project is about the characterization of a single domain. Each group is provided with a representative **domain sequence** and the corresponding **Pfam identifier** (see table below). The objective of the project is to build a sequence **model** starting from the assigned sequence and to provide a **functional characterization** of the entire domain family (homologous proteins).

The analysis of the results will be delivered in a **PDF report** of at least two and not more than five pages of text, excluding figures and supporting documentation. Domain models, code, commands and generated data will be delivered as **supplementary material** (compressed archive). Clarity of the documentation and the reproducibility of the analysis will be evaluated along with the performance of the models which should be comparable to the corresponding Pfam. The project has to be submitted at least **10 days before** the exam date to **damiano.piovesan@unipd.it**.

Input

A representative sequence of the domain family. Columns are: group, UniProt accession, organism, Pfam identifier, Pfam name, domain position in the corresponding UniProt protein, domain sequence. **Group assignments** are provided **here** (students sheet).

Domain model definition

The objective of the first part of the project is to build a **PSSM** and **HMM** model representing the assigned domain. The two models will be generated starting from the assigned **input sequence**. The accuracy of the models will be evaluated against **Pfam** annotations as provided in the SwissProt database.

Models building

1. Retrieve homologous proteins starting from your input sequence performing a **BLAST search** against UniProt or UniRef50 or UniRef90, or any other database
2. Generate a **multiple sequence alignment (MSA)** starting from retrieved hits using T-coffee or ClustalOmega or MUSCLE
3. If necessary, edit the MSA with JalView (or with your custom script or CD-HIT) to **remove not conserved positions** (columns) and/or **redundant information** (rows)
4. Build a **PSSM** model starting from the MSA

5. Build a **HMM** model starting from the MSA

Models evaluation

1. Generate **predictions**. Run **HMM-SEARCH** and **PSI-BLAST** with your models against SwissProt.
    a. Collect the list of retrieved hits
    b. Collect matching positions of your models in the retrieved hits
2. Define your **ground truth.** Find all proteins in SwissProt annotated (and not annotated) with the assigned Pfam domain
    a. Collect the list of proteins matching the assigned Pfam domain
    b. Collect matching positions of the Pfam domain in the retrieved sequences. Domain positions are available here (large tsv file) or using the InterPro API or align the Pfam domain yourself against SwissProt (HMMSEARCH)
3. **Compare** your model with the assigned Pfam. Calculate the precision, recall, F-score, balanced accuracy, MCC
    a. Comparison at the **protein level**. Measure the ability of your model to retrieve the same proteins matched by Pfam
    b. Comparison at the **residue level**. Measure the ability of your model to match the same position matched by Pfam
4. Consider refining your models to **improve their performance**

Domain family characterization

Once the family model is defined (previous step), you will look at functional (and structural) aspects/properties of the entire protein family. The objective is to provide insights about the main function of the family.

Taxonomy

1. Collect the **taxonomic lineage** (tree branch) for each protein of the *family_sequences* dataset from UniProt (entity/organism/lineage in the UniProt XML)
2. Plot the **taxonomic tree** of the family with nodes size proportional to their relative abundance

Function

1. Collect **GO annotations** for each protein of the *family_sequences* dataset (entity/dbReference type="GO" in the UniProt XML)
2. Calculate the **enrichment** of each term in the dataset compared to GO annotations available in the SwissProt database (you can download the entire SwissProt XML here). You can use Fisher'

exact test and verify that both two-tails and right-tail P-values (or left-tail depending on how you build the confusion matrix) are close to zero

3. Plot enriched terms in a **word cloud**
4. Take into consideration the hierarchical structure of the GO ontology and report most significantly enriched **branches**, i.e. high level terms
5. Always report the **full name** of the terms and not only the GO ID

Motifs

1. Search significantly conserved short **motifs** inside your family. Use **ELM classes** and **ProSite patterns** (for ProSite consider only patterns "PA" lines, not the profiles). Make sure to consider as true matches only those that are found inside disordered regions. Disordered regions for the entire SwissProt (as defined by MobiDB-lite) are available here

Useful Software

- JalView (http://www.jalview.org). Multiple sequence alignment viewer.Clustal-Omega. (http://www.clustal.org/omega/). Multiple sequence alignment.
- HMMER (http://hmmer.org/). Build HMM models of multiple sequence alignments. Perform HMM/sequence database searches.
- NCBI-BLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Perform database sequence searches.

Useful databases

- UniProt, https://www.uniprot.org/
- PDB, https://www.rcsb.org/
- InterPro, https://www.ebi.ac.uk/interpro/
- Pfam, https://pfam.xfam.org/
- Gene Ontology, http://geneontology.org/docs/download-ontology/