Data Science

Biological Data Project - 2020-21

(Version - 22/12/2020)

*"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.**

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.

Building the models:

1. Define your **ground truth** by finding all proteins in SwissProt annotated (and not annotated) with the assigned Pfam domain and collect the position of the Pfam domain for all sequences. Domain positions are available here or using the InterPro API.
2. Retrieve homologous proteins starting from your input sequence performing a **BLAST search** against UniProt or UniRef50 or UniRef90.
3. Generate a **multiple sequence alignment (MSA)** starting from retrieved hits using T-coffee or ClustalOmega or MUSCLE.
4. If necessary, edit the MSA with JalView (or with your custom script) to **remove noise**.
5. Build a **PSSM** model starting from the MSA.
6. Build a **HMM** model starting from the MSA.
7. Find significant hits using **HMM-SEARCH** and **PSI-BLAST** against SwissProt.
8. **Evaluate** the ability of matching **sequences** considering your ground truth. Calculate accuracy, precision, sensitivity, specificity, MCC, F-score, etc.
9. **Evaluate** the ability of matching domain **position** considering your ground truth, i.e. residues overlapping (and non overlapping) with Pfam domains. Calculate accuracy, precision, sensitivity, specificity, MCC, F-score, etc.
10. Consider repeating point 2-4 to **improve the performance** of your models.

11. Choose the best model.

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.

Dataset definitions:

- **family_structures** - All PDB chains whose sequences significantly match your model and with a minimum overlap of 80%. If necessary, e.g. if you get more than 50 PDB chains, reduce the size of *family_structures* clustering by sequence identity.
- **family_sequences** - All UniRef90 sequences matching your model. Limit your result to max 1,000 proteins. UniProt annotation (entries XML files) can be retrieved with the "Retrieve/ID mapping" service" from the UniProt website.

Structural characterization

1. Perform an **all-vs-all pairwise structural alignment** using the TM-align software.
2. Build a matrix representing the **pairwise RMSD** and/or the TM-score provided by TM-align in the previous step for all possible pairs of structures.
3. Calculate a **dendrogram** representing a hierarchical clustering of the matrix. You can use *scipy.cluster.hierarchy.linkage* and *scipy.cluster.hierarchy.dendrogram* Python methods.
4. Remove outliers.
5. Identify conserved positions performing a **multiple structural alignment** of the *family_structures* dataset.
6. Identify long range (sequence separation ≥ 12) **conserved contacts**. You can align the contact maps of each structure based on the multiple structural alignment and identify conserved positions.
7. Identify the **CATH** superfamily (or superfamilies) and family (or families) matching your model, if any.

Taxonomy

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

Functional characterization

10. Collect **GO annotations** for each protein of the *family_sequences* dataset (entity/dbReference type="GO" in the UniProt XML).
11. 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.
12. Plot enriched terms in a **word cloud**.
13. Take into consideration the hierarchical structure of the GO ontology and report most significantly enriched **branches**, i.e. high level terms.

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.
- TM-align (https://zhanglab.ccmb.med.umich.edu/TM-align/). Perform pairwise structural alignments.

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/