Project

Data Science - Biological Data - 2019/2020

(Version - 27/11/2019)

"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) in human. An introduction about the function of the assigned domains is available in this book.

The analysis of the results will be delivered in a 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 biocomp@bio.unipd.it.

Group #	Domain name	Domain (Pfam)	UniProt (start-end)	Sequence
1, 8	SH2	SH2 (PF00017)	P23615 (1258-1339)	YYFPFNGRQAEDYLRSKERGEFVIRQSSRGDDHLVITWKLDKDLFQHIDIQELEKENPLALGKVLIVDNQKYNDLDQIIVEY
2, 7	SH3	SH3_1 (PF00018)	P36006 (1126-1174)	EAAYDFPGSGSSSELPLKKGDIVFISRDEPSGWSLAKLLDGSKEGWVPT
3, 6	WW	WW (PF00397)	P43582 (11-41)	VPSGWKAVFDDEYQTWYYVDLSTNSSQWEPP
4, 5	PDZ	PDZ (PF00595)	P53920 (290-375)	QWLLKPYDECRRLGLTSERESEARAKFPENIGLLVAETVLREGPGYDKIKEGDTLISINGETISSFMQVDKIQDENVGKEIQLVIQ

Domain models

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 from the input sequence (provided) and homologous sequences retrieved from UniProt (not provided). The accuracy of the models will be evaluated against Pfam annotations of the corresponding domain (Pfam ID provided) in the Human organism.

Building the models:

1. Retrieve homologous proteins in UniProt starting from the input sequence provided.
2. Generate a multiple sequence alignment (MSA). If necessary, (manually) edit rows and columns.
3. Build a PSSM model starting from the MSA using BLAST.
4. Build a HMM model starting from the MSA using HMMER.
5. Evaluate your model against human proteins available in SwissProt (accuracy, precision, sensitivity, specificity, MCC):

a. Define you ground truth/reference by finding all human proteins in SwissProt annotated (and not annotated) with the assigned Pfam ID (provided). Pfam annotations are available from UniProt.

b. Find significant hits using HMM-SEARCH and PSI-BLAST respectively for the HMM and PSSM model.

c. Evaluate the ability of retrieving proteins with that domain.

d. Evaluate the ability of matching the domain position, i.e. the alignment position of the model in the retrieved proteins (Pfam reference position is available in InterPro).

6. Consider repeating point 1-4 to improve the performance of your models.

Domain family characterization

Once the list of human proteins matching your models is defined (previous step), you will look at functional and structural aspects/properties. First, you will compare you dataset with the entire human proteome. Second, you will split the dataset considering the different domain architectures (other Pfam domains) and compare with the original set. Third, you will expand the original set by including direct interactors from protein-protein interaction network and compare its properties with the original dataset.

Dataset definitions:

- Original dataset. The proteins retrieved with your model against human proteins in SwissProt.
- Architectures datasets. Each dataset contains proteins with the same combination of Pfam domains. The order of the domains in the sequence and repetition are not important.
- PDB network. The proteins of the original dataset with a PDB plus other human proteins which are found as other chains in the same PDB.
- STRING network. The proteins of the original dataset plus all direct interactors found in the STRING database.

Properties to calculate for each dataset:

- Annotation enrichment (Gene Ontology - GO and Disease Ontology - DO)
  - Measure enrichment performing a Fisher’s exact test (hypergeometric test).
  - Take into consideration the hierarchical structure of the ontologies and report only most enriched branches, i.e. high level terms.
  - Report enriched terms in a “word cloud” representation (also including specific terms).
  - For each dataset the background is:
    - Original dataset - The entire human proteome available in SwissProt.
    - Architectures datasets - The original dataset.
    - PDB network - All human PDB proteins available in SwissProt (intersection).
    - STRING network - All human proteins available in STRING and SwissProt (intersection).
  - Annotations sources:
    - Retrieve GO annotation from UniProt GOA
    - Mine DO terms from those articles (abstracts) associated with dataset proteins in EuropePMC.
  - Ontologies:
    - GO OBO file here.
    - DO OBO file here.
- Structural classification
  - Provide a statistics about the CATH architectures mapping to your domain.
  - Retrieve all PDBs covering your domain (if any) and evaluate their structural similarity.
    - Perform an all-vs-all pairwise structural alignment using the TM-align software.

        - Build a distance matrix considering the RMSD and/or the TM-score provided by TM-align in the previous step for all possible pairs of structures.
        - Calculate a dendrogram representing a hierarchical clustering of the matrix (you can use "scipy.cluster.hierarchy.linkage" and "scipy.cluster.hierarchy.dendrogram" Python functions).

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/
- STRING, https://string-db.org/cgi/input.pl
- GOA, https://www.ebi.ac.uk/GOA/