|  |  |  |
| --- | --- | --- |
| *Name* | *Surname* | *ID* |

## Midterm test No. 2

### 24 / 11 / 2020

Please answer all questions below and submit this document in **PDF format** by **12:30 of the 1st December 2020** (one week after) to **damiano.piovesan@unipd.it**.

Each student is assigned a different **CATH superfamily** and a set of 10 representative domains. The entire exercise is based on the analysis of that superfamily. For each question **concisely explain all passages** **(max 5 rows)** necessary to reproduce the results (e.g. parameters, database queries, algorithms, etc.). Optionally, if relevant, you can provide source code (not necessary).

Superfamily student **assignments** are available[**here**](https://docs.google.com/spreadsheets/d/15s6AtfeArnmyBqLp0jWKMMb7rIpo8r51-ZvVlIUT10k/edit?usp=sharing).

Superfamily **representative domains** are available [**here**](https://drive.google.com/file/d/1Dvj_AYeq7NoN-BIz2n4fwrv22lEfzi1D/view?usp=sharing). (Columns: PDB ID, chain ID, PDB domain start, PDB domain end, domain sequence)

## Questions

1. Paste below your assigned CATH superfamily identifier.
2. Compare the sequences of your superfamily provided in the assignment file performing an all-vs-all pairwise sequence alignment.
   1. Paste below a 10 x 10 matrix where cells represent the pairwise sequence identity.
   2. Which is the domain more similar to all other domains?
   3. Based on sequence identity (e.g. 30% threshold), are there domains which can be grouped in the same family?
3. Download the PDB files associated with your CATH superfamily and answer the following questions considering the start/end positions of the domain fragment as provided in the assignment file.
   1. Which is the coverage of your domain fragments on the corresponding PDB chains (consider observed residues)?
   2. Which is the coverage of your domain fragments on the corresponding full length proteins (UniProt sequences)?
4. For each PDB create a new PDB with the coordinates of the domain fragment and perform an all-vs-all pairwise structural alignment using TM-align.
   1. Paste below a 10 x 10 matrix where cells represent the pairwise sequence identity obtained with the structural alignment (not sequence alignment).
   2. Paste below a 10 x 10 matrix where cells represent the pairwise RMSD.
   3. Which is the domain more similar to all other domains looking at the sequence identity (calculated with the structural alignment)?
   4. Which is the domain more similar to all other domains looking at the RMSD?
5. Create a multiple sequence alignment (MSA) starting from the domain sequences available in the assignment file using EBI T-Coffee.
   1. Which are the most conserved columns looking at the amino acid composition?
   2. Which are the most conserved columns looking at the column entropy?
6. Use the MSA generated before to perform a PSI-BLAST and a HMMER search against human proteins.
   1. How many significant hits are returned by the two methods?
7. Which PFAM HMMs match your superfamily? **Hint**: you can use hmmscan EBI service.