Biological Data
University of Padova - Department of Mathematics
26th February 2021
Piero Romare, Giorgio Dilda, Gleydson Da Silva, Veniamin Pavlov

Project

Our assignment:

- Group: 2
- Uniprot: P07897
- Organism: Rattus norvegicus (Rat)
- PFAM ID: PF00193
- PFAM Name: Link domain
- Domain position: 254-349
- Domain sequence: EVFYATSPEKFTFQEAANECRTVGARLATTGQLYLAWQG-GMDMCSAGWL ADRSVRYPISKARPNCGGNLLGVRTVYLHANQTGYPDPSS-RYDAICY

Domain Models

Our objective is to build a PSSM and HMM models representing the domain and subsequently the evaluation of the models against Pfam annotations of our domain. First of all, we retrieved the homologous proteins in Uniprot by searching our input sequence (in Uniprot identified by id P07897, with our domain located from residue 254 to residue 349). We performed a BLAST search in Uniprot website with different calibration of parameters: https://www.ebi.ac.uk/Tools/sss/ncbiblast/

- UniprotKB, e-value 0.001, 500 hits, BLOSUM62 matrix;
- Uniref90, e-value 0.001, 500 hits, BLOSUM62 matrix;
- UniRef50, e-value 0.001, 500 hits, BLOSUM62 matrix.

We want to highlight the fact that our domain can be found in many other proteins and they can be multi-domain, meaning that our domain would be included in them, but also other domains that are not the characterisation of this work. After the retrieval process, we generated a Multiple Sequence Alignment (MSA) from our BLAST search result using the Clustal Omega https://www.ebi.ac.uk/Tools/msa/clustalo/ , and we used Jalview to visualize the alignment.
After that, starting from the MSAs, we

- build a PSSM from the MSA using PSI-BLAST;
- build a HMM from the MSA with HMM-SEARCH.

At this point, we searched with our PSSM and HMM models against the database SwissProt using HMM-SEARCH and PSI-BLAST. Now, the next step consists on evaluating and comparing the ability of matching sequences of our two models against proteins available in SwissProt. So, we downloaded dataset wrt our assigned PFAM ID PF00193 from Uniprot with the following (pf00193 AND reviewed:yes). The number of proteins found on SwissProt are 58:

- PSSM coming from Uniref90's MSA
  - accuracy: 0.997
  - precision: 1.0
  - sensitivity: 0.773
  - specificity: 1.0
  - mcc: 0.878
  - f1: 0.87
- For what concern HMM, best ones is HMM of MSA coming from Uniref90, here following HMM metrics:
  - accuracy: 0.997
  - precision: 1.0
  - sensitivity: 0.773
  - specificity: 1.0
  - mcc: 0.87
  - f1: 0.872

Now, the next step consists on evaluating and comparing the ability of matching domain position of our two models against our ground truth. Thus, we downloaded dataset w.r.t our link domain PFAM ID:

- PSSM
  - avg acc: 0.977,
  - avg precision: 0.997,
  - avg recall: 0.998,
  - avg specificity: 0.976,
  - avg mcc: 0.934,
  - avg f1: 0.949
- HMM
  - avg acc: 0.983
  - avg precision: 0.954,
  - avg recall: 0.979
  - avg specificity: 0.98
  - avg mcc: 0.952,
  - avg f1: 0.966

Domain family characterization

Dataset

By using HMM of from MSA of Blast search on Uniref90 model, we define *family structures* and *family sequences*. For what concern structures, we retrieve 13 PDB chains whose sequences match by performing hmmsearch online tool against PDB database. For the second one we retrieve over 7.000 UniRef90 sequences matching our model. We order our sequences in decreasing order of e-value, and we select the first 1.000. Due to the fact that sequences which have start with "UPI" are not found (mapped in Uniprot KB), following the order previously described, we select the following sequences which not start with "UPI".

Structural Characterization

Unable to install TM-Align, we run it in a VM Ubuntu OS and perform all-vs-all pairwise strucutal alignment using TM-Align, following the dataframe described scores, RMSD and Dendogram (Fig. 1, 2, 3):

pdb1o7b	pdb1poz	pdb2i83	pdb2jcp	pdb2jcq	pdb4mrd	pdb4mre	pdb4pz3	pdb4pz4	pdb5bzc	pdb5bze	pdb5bzf	pdb5xts
1.00000	0.66625	0.74171	0.76842	0.77044	0.77592	0.77147	0.78354	0.77393	0.77059	0.76978	0.77086	0.60144
0.45028	1.00000	0.62022	0.74305	0.74167	0.74121	0.74257	0.72246	0.73797	0.74073	0.74180	0.73868	0.40214
0.49651	0.62333	1.00000	0.71237	0.71361	0.71259	0.71250	0.72282	0.72309	0.71162	0.71078	0.70827	0.43118
0.53212	0.78028	0.74501	1.00000	0.99210	0.98729	0.99787	0.96689	0.97823	0.99776	0.99867	0.99565	0.45296
0.53628	0.78344	0.75071	0.99875	1.00000	0.99451	0.99775	0.96793	0.97814	0.99715	0.99811	0.99468	0.44435
0.53938	0.78282	0.74959	0.99388	0.99451	1.00000	0.99192	0.96465	0.97612	0.99200	0.99301	0.99091	0.45250
0.53363	0.77980	0.74507	0.99787	0.99111	0.98535	1.00000	0.96536	0.97503	0.99883	0.99897	0.99667	0.45172
0.54016	0.75782	0.75638	0.96689	0.96162	0.95838	0.96536	1.00000	0.98226	0.96620	0.96654	0.96654	0.45606
0.52234	0.75787	0.73953	0.95318	0.94684	0.94492	0.95014	0.95712	1.00000	0.95033	0.95133	0.94965	0.45019
0.53329	0.77760	0.74412	0.99776	0.99053	0.98543	0.99883	0.96620	0.97523	1.00000	0.99942	0.99806	0.45140
0.53278	0.77908	0.74325	0.99867	0.99147	0.98643	0.99897	0.96654	0.97628	0.99942	1.00000	0.99760	0.45371
0.53353	0.77554	0.74055	0.99565	0.98809	0.98436	0.99667	0.96654	0.97451	0.99806	0.99760	1.00000	0.45197
0.15196	0.15978	0.16726	0.16750	0.16104	0.16578	0.16789	0.16790	0.17064	0.16790	0.16758	0.16829	1.00000

Figure 1: All-vs-All Pairwise TM-Score

	1o7b	1poz	2i83	2jcp	2jcq	4mrd	4mre	4pz3	4pz4	5bzc	5bze	5bzf	5xts
1o7b	0.000	2.847	2.193	2.436	2.428	2.384	2.418	2.118	2.192	2.407	2.422	2.386	2.451
1poz	2.847	0.000	3.506	2.842	2.794	2.793	2.719	3.063	3.035	2.752	2.744	2.769	3.502
2i83	2.193	3.506	0.000	2.677	2.662	2.669	2.797	2.524	2.545	2.809	2.694	2.696	2.886
2jcp	2.436	2.842	2.677	0.000	0.161	0.384	0.213	0.833	0.589	0.217	0.167	0.303	3.158
2jcq	2.428	2.794	2.662	0.161	0.000	0.367	0.217	0.815	0.588	0.244	0.198	0.334	3.119
4mrd	2.384	2.793	2.669	0.384	0.367	0.000	0.452	0.873	0.639	0.450	0.421	0.475	3.117
4mre	2.418	2.719	2.797	0.213	0.217	0.452	0.000	0.856	0.649	0.157	0.147	0.264	3.083
4pz3	2.118	3.063	2.524	0.833	0.815	0.873	0.856	0.000	0.707	0.845	0.842	0.834	2.891
4pz4	2.192	3.035	2.545	0.589	0.588	0.639	0.649	0.707	0.000	0.648	0.627	0.656	3.215
5bzc	2.407	2.752	2.809	0.217	0.244	0.450	0.157	0.845	0.648	0.000	0.110	0.201	3.079
5bze	2.422	2.744	2.694	0.167	0.198	0.421	0.147	0.842	0.627	0.110	0.000	0.224	3.150
5bzf	2.386	2.769	2.696	0.303	0.334	0.475	0.264	0.834	0.656	0.201	0.224	0.000	3.250
5xts	2.451	3.502	2.886	3.158	3.119	3.117	3.083	2.891	3.215	3.079	3.150	3.250	0.000

Figure 2: RMSD

By using MultiTM-Aligh online tool, we perform multiple structural alignment.

https://zhanglab.ccmb.med.umich.edu/TM-align/ .

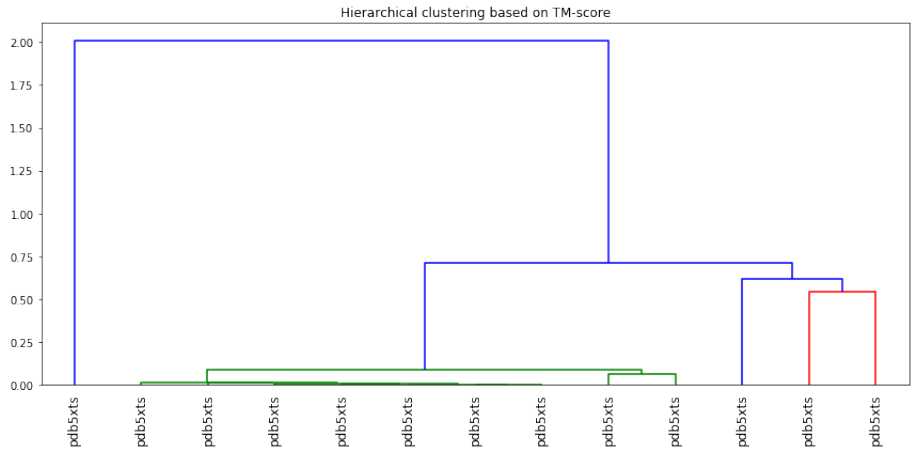

Figure 3: Dendogram All-vs-All Pairwise TM-Score

We identify 90 long range conserved contacts with sequence separation greater or equal to 12 and 63 with sequence separation greater or equal to 13. In order to be clear, it follows first 5 results (entropy, non gap, columns):

- 0.09 13 ['G' 'G' 'G' 'G' 'G' 'G' 'G' 'G' 'G' 'G' 'G' 'G' 'S']
- 0.09 13 ['V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'Y']
- 0.18 13 ['Y' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'Q']
- 0.09 13 ['H' 'H' 'H' 'H' 'H' 'H' 'H' 'H' 'H' 'H' 'H' 'H' 'I']
- 0.18 13 ['R' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'V' 'N']

Taxonomy

From Uniprot we collect the taxonomy. We obtain a total of 629 species. To the best of our library's knowledge, we plot the tree with high resolution to obtain little readability (Fig. 4). We'd provide an example of dict in which, by opening the script, it can be interacted:
{'Eukaryota': {'Metazoa': {'Chordata': {'Craniata': {'Vertebrata': {'Euteleostomi': {'Actinopterygii': {'Chondrostei': {'Acipenseriformes': {'Acipenseridae': {'Acipenser': {}}}}}, 'Neopterygii': {'Teleostei': {'Neoteleostei': {'Acanthomorphata': {'Carangaria': {'Carangaria incertae sedis': {'Centropomidae': {'Lates': {}}},
Another alternative is to take in consideration a Counter format (parent, child) as

1. ('Eukaryota', 'Metazoa'): 990,
2. ('Metazoa', 'Chordata'): 990,
3. ('Chordata', 'Craniata'): 987,
4. ('Craniata', 'Vertebrata'): 987,
5. ('Vertebrata', 'Euteleostomi'): 974,
6. ('Euteleostomi', 'Actinopterygii'): 488,

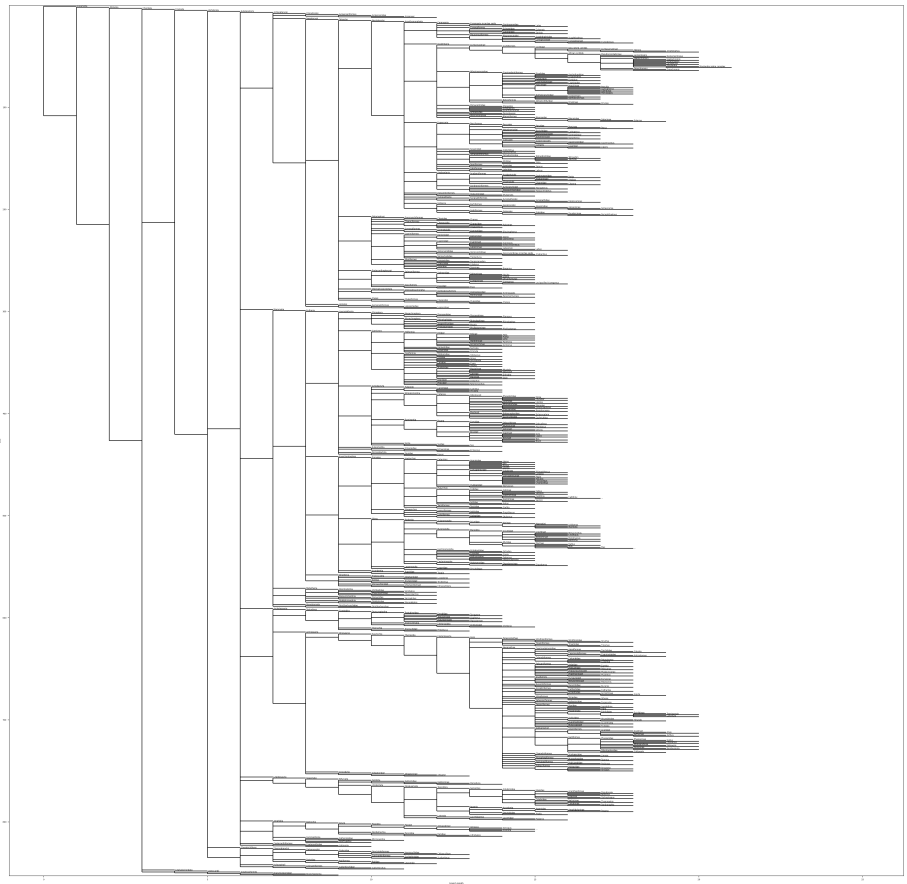

Figure 4: Taxonomy Tree

7. ('Actinopterygii', 'Chondrostei'): 1,
8. ('Chondrostei', 'Acipenseriformes'): 1,
9. ('Acipenseriformes', 'Acipenseridae'): 1,
10. ('Acipenseridae', 'Acipenser'): 1,
11. ('Euteleostomi', 'Mammalia'): 313
12. ...

Functional Characterization

From Uniprot, we collect GO annotations for each protein in our *family sequences* dataset. By using egreb command function we are able to parse the entire Swissprot XML in order to retrieve accession and annotations and to perform Fisher test for each term. We plot the WordCloud (Fig. 5). and here we report the most significantly enriched branches

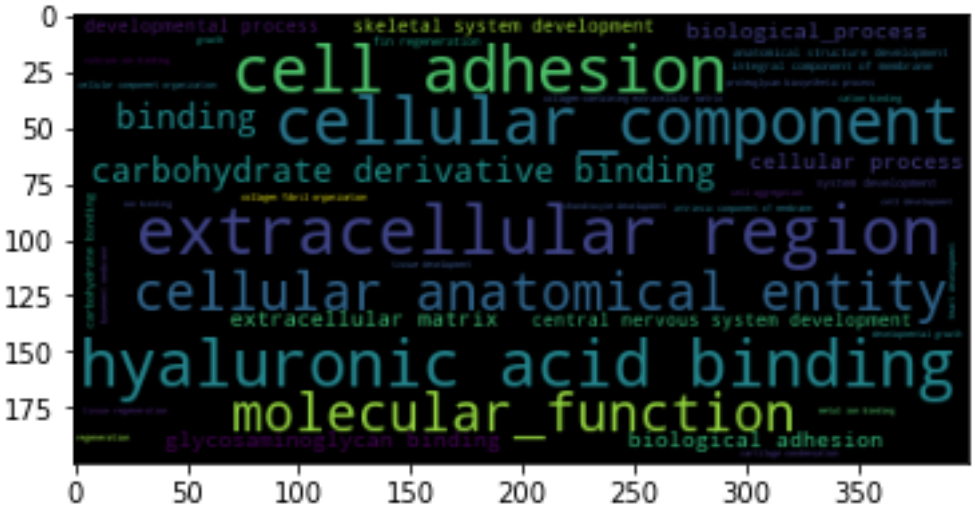

Figure 5: WordCloud GO

(annotation - count - sp_count, enrichment - depth)

1. GO:0005576 - 717 - 26453 - 0.0 - 2.0;
2. GO:0005509 - 374 - 4771 - 0.0 - 5.0;
3. GO:0022610 - 877 - 2448 - 0.0 - 1.0;
4. GO:0009987 - 879 - 64943 - 0.0 - 1.0;
5. GO:0005539 - 877 - 177 - 0.0 - 3.0.

Appendix

In order to be clear, we create a main folder, two subfolder named *part_ 1* and *part_ 2*, respectively models part and functional and structural aspects part. For each subfolder we divide other subfolders, in most cases, related to each step of your guidelines.

Part 1

1. The output of **BLAST** search is stored in 50.fasta, 90.fasta and kb.fasta.
2. The output of **Clustal Omega** alignment is stored in clustalo-50.clustal, clustalo-90.clustal, clustalo-kb.clustal
3. To create the PSSM we used the command **psiblast** with the default parameters, passing the MSA file and the Swissprot database which produced two files, a .pssm model and .txt result from the search.
4. To create the HMM, first we run the command **hmmbuild** in which expected two parameters, the model to be generated from the MSA file and the MSA itself. Then, from the model we performed an **hmmsearch** against the Swissprot database producing three different output formats, tblout, domtblout, align.
5. uniprot-pf00193-reviewed.fasta containing our proteins domain in Uniprot.
6. PF00193.json containing our domain position in InterPRO.

Part 2

1. We create *family structures* using **HMM-SEARCH** against PDB database obtaining a list stored in hmmsearch_hmm90_pdb.tsv All PDBs downloaded in dataset-pdb folder
2. We create *family sequences* in 990.xml file
3. All-vs-all pairwise is stored in tm-output.txt file
4. Multiple structural alignment score is stored in TMscore.txt
5. dict.pkl contains a dictionary data structure of tree
6. parent_children.pkl contains a list of tuple data structure of tree
7. go_swissprot.txt contains a parse version of entire swissprot xml