Family Model

Final Project

Biological Data

Team Members
Bogdana Živković
Nikola Ivanović
Prashant Gupta
Zyad Al-Azazi

Course Instructors:
Damiano Piovesan
Silvio Tosatto

Models Building:

The first step in building the family models was retrieving the homologous proteins using the EBI BLAST web service. The search was performed on the UniProt database with Alignments parameter set to 1000 and Alignment Views parameter set to BLASTXML. The output was downloaded as an XML file, which was then parsed in order to extract hit IDs. Then, using the UniProtKB Retrieve/ID mapping tool the hit IDs were mapped to sequences in fasta format. The next step was computing the MSA using the Clustal Omega web service by providing the previously obtained sequences. Additionally, the MSA file was edited in order to test performance in successive steps of the analysis. Using JalView, unconserved columns in the MSA (i.e. columns that have '-' in the conservation histogram) were manually removed and redundant rows were filtered out. By keeping the redundancy threshold at 100, only the identical sequences were removed. Later steps were performed with both the original and the edited MSA. The PSSM and the HMM were built using the PSI-BLAST command line tool and the HMM command line tool respectively.

Models Evaluation:

In order to perform model evaluation it was necessary to obtain protein family predictions using the created models. For the PSSM, NCBI BLAST was used for predictions, making sure to select Swiss-Prot as the database and selecting PSI-BLAST as the program. In the case of HMM, HMMSEARCH was used for the same task, selecting SwissProt as the sequence database. The outputs for both tools were downloaded for later evaluation. Using the Pfam domain ID (PF08534), accession search was performed on SwissProt with HMMSEARCH and the results were downloaded. Considering that HMMSEARCH only provides UniProt IDs and not Accessions, they had to be mapped with the UniProt mapping service in order to evaluate the PSSM performance. It should be noted that the evaluations were run for both models constructed from the original and edited MSA, as well as for multiple iterations for PSSM models.

Both at protein level and residue level, it was necessary to compute true positives, false positives, false negatives and true negatives in order to obtain the final performance metrics. At the protein level TP corresponds to the number of common sequences between the model and the Pfam domain, FP to the number of proteins only present in the model, FN to the number of proteins only present in the Pfam domain and TN to the number of proteins in SwissProt that are not present neither in the model or the Pfam domain. The results of the analysis indicate that the models obtained from the edited MSA perform slightly better and the later iterations for PSSM perform noticeably better, which is reflected in scores other than precision, as they take into account the number of false negatives.

File	TP	FP	FN	Precision	Recall	F1	Balanced Accuracy	MCC
HMM	32	6	418	0.842105	0.071111	0.131148	0.535550	0.244586

File	TP	FP	FN	Precision	Recall	F1	Balanced Accuracy	MCC
HMM for conserved	47	0	403	1.000000	0.104444	0.189135	0.552222	0.323064
PSSM	32	0	418	1.000000	0.071111	0.132780	0.535556	0.266569
Conserved PSSM - Iteration 1	33	0	417	1.000000	0.073333	0.136646	0.536667	0.270702
Conserved PSSM - Iteration 2	145	2	305	0.986395	0.322222	0.485762	0.661109	0.563616
Conserved PSSM - Iteration 3	233	5	217	0.978992	0.517778	0.677326	0.758885	0.711825
Conserved PSSM - Iteration 4	249	10	201	0.961390	0.553333	0.702398	0.776658	0.729217
Conserved PSSM - Iteration 5	254	13	196	0.951311	0.564444	0.708508	0.782211	0.732630
Conserved PSSM - Iteration 6	264	18	186	0.936170	0.586667	0.721311	0.793318	0.740945
Conserved PSSM - Iteration 7	270	22	180	0.924658	0.600000	0.727763	0.799981	0.744694
Unconserved PSSM - Iteration 2	100	0	350	1.000000	0.222222	0.363636	0.611111	0.471260
Unconserved PSSM - Iteration 3	98	2	352	0.980000	0.217778	0.356364	0.608887	0.461827
Unconserved PSSM - Iteration 4	97	3	353	0.970000	0.215556	0.352727	0.607775	0.457111
Unconserved PSSM - Iteration 5	97	3	353	0.970000	0.215556	0.352727	0.607775	0.457111

Table 1. Performance Metrics for Sequence Alignment Models (Protein Level)

At the residue level, only matched positions of the corresponding sequences were compared. That is, we compared target alignment positions in the model with the ones in the Pfam domain for matching sequences. In this case, TP represents the count of correctly predicted residue positions, FP the count of predicted residue positions that were not in the Pfam domain, FN the count of matched positions in the Pfam domain that were not predicted by the model and TN the count of residue positions not matched by either. As in the previous case, precision, recall, F1, balanced accuracy and MCC were computed from the confusion matrix. It can be concluded from the results that the differences in performance are not as drastic as at the protein level, with the higher iterations of PSSM models performing slightly worse in most metrics. Finally, considering the results in total, we can conclude that the model's ability to accurately include proteins present in the family comes at the cost of reduced performance at predicting residue positions. When everything is taken into account we deemed the "Conserved PSSM - Iteration 7" model the best representation of the protein family and used it for further analysis.

File	TP	FP	FN	TN	Precision	Recall	F1	Balanced Accuracy	MCC
HMM	4157	199	598	1507	0.954316	0.874238	0.912523	0.878795	0.712573
HMM for conserved	5635	124	1191	2413	0.978468	0.825520	0.895511	0.888322	0.709413
PSSM	4766	336	256	1570	0.934143	0.949024	0.941525	0.886369	0.783271
Conserved PSSM - Iteration 1	4654	270	378	1626	0.945167	0.924881	0.934914	0.891238	0.769400
Conserved PSSM - Iteration 2	21541	2099	1705	8848	0.911210	0.926654	0.918867	0.867456	0.742242
Conserved PSSM - Iteration 3	32165	7337	312	7011	0.814263	0.990393	0.893733	0.739516	0.607988
Conserved PSSM - Iteration 4	34056	9911	95	5352	0.774581	0.997218	0.871912	0.673935	0.513212
Conserved PSSM - Iteration 5	34951	11398	212	7139	0.754083	0.993971	0.857567	0.689546	0.524342
Conserved PSSM - Iteration 6	35626	12686	740	10397	0.737415	0.979651	0.841446	0.715035	0.537181
Conserved PSSM - Iteration 7	36815	14832	1484	14946	0.712820	0.961252	0.818602	0.731583	0.536954
Unconserved PSSM - Iteration 2	13391	1986	662	4824	0.870846	0.952893	0.910024	0.830631	0.704317
Unconserved PSSM - Iteration 3	13845	4033	18	3050	0.774415	0.998702	0.872373	0.714655	0.574409
Unconserved PSSM - Iteration 4	13533	4562	9	2940	0.747886	0.999335	0.855517	0.695615	0.539817
Unconserved PSSM - Iteration 5	13475	4674	4	2829	0.742465	0.999703	0.852093	0.688376	0.528388

Table 2. Performance Metrics for Sequence Alignment Models (Residue Level)

Domain Family Characterization:

The protein family explored in this project is the Peroxiredoxin, AhpC-type protein family. This protein family is part of another family called Prx1 subfamily from the whole Peroxiredoxins (Prxs) family. The Prx1 subfamily is widespread across all of the animal kingdom. The family of Prxs is regarded to be very important in antioxidant protection and cellular pathways. The general structure of the Prxs family is a globular protein structure that includes a Thioredoxin (Trx) fold, which is the superfamily/ clan that the Prxs descend from. The members of this family also host four well conserved residues among which is the active site Cysteine (Cys) or peroxidatic Cys (CP); a residue that is significantly important in the catalytic cycle that reduces the peroxide substrates - Catalysis - by utilizing the reaction of the CP with the peroxide substrate in Prxs. The catalytic cycle is constituted by three stages:

peroxidation, resolution and recycling. The importance of Prxs is in their role as cellular antioxidant enzymes (in some organisms, the only antioxidants) effective against Reactive Oxygen Species (ROS) that are capable of damaging biological macromolecules through their relations to diseases, such as neurodegenerative disease and cancer.

Taxonomy:

The following steps explain in detail how the taxonomy lineages of the proteins of the family were obtained:

- After selecting the best performing model, the Uniprot IDs of the proteins were retrieved from the PSSM model hit table.
- The Uniprot IDs were used to retrieve the corresponding UniRef90 IDs from here.
- The UniRef90 IDs were used to retrieve the proteins and their taxonomy IDs from the UniRef90 database by exporting the results into a JSON file.
- To obtain the taxonomy lineage of each protein individually, API calls to the Entrez database using the taxon IDs were used.

The next step was to create the taxonomic tree of all the family proteins with the following steps:

- First, the unique taxon IDs were extracted to a text file.
- Second, the Common Taxonomy Tree tool was used to obtain the .phy file used to plot the taxonomic tree.
- Finally, the plot for the taxonomic tree was created on a local machine due to some unresolved error within the ete3 package in Google Colab.

The final taxonomic tree can be found here.

Function:

For the Protein functionality we have to identify the Go annotation. For this we have used the xml file.

Note - As Xml file is huge and it was not possible to process it with python, we have perform below steps to process and get mapping between Uniprot Id to Go annotation.

1. **Download UniProt XML File:** Download the `uniprot_sprot.xml` file from UniProt Downloads.
2. **Preprocess XML using Grep:** Open a terminal and use the following command to process the XML file, extracting relevant information into a temporary file named `processed_sprot.txt`.

```
grep -E '<entry|<accession|<dbReference type="GO" id="GO:'
uniprot_sprot.xml > processed_sprot.txt
```

3. Python code is used to create from `processed_sprot.txt`.

We have used the uniprot ids of the best model to get the children and ancestor annotation with the help of API.

In the dataset, we were able to observe the results in *Table 3*:

	Family Proteins	**Family Proteins with Lineage**
Number of GOAs	1582	10437
Number of Unique GOAs	200	801

Table3: The total number of GO annotations and unique GO annotations in the family proteins and along with their taxonomic lineage.

After this, we created a confusion matrix for each GO annotation as demonstrated in *Table 4*. Then, we applied the two-sided, left-sided and right-sided Fisher exact tests to obtain the significance. We also applied the fold increase technique to attain the enrichment of the GO terms. We only considered GO terms with enrichment values greater than 1 and with two-sided significance levels of 0.01 or less.

GO:0043229	**Property Present**	**Property Not Present**
Selected	63	231
Not Selected	216991	568511

Table 4: An example of the confusion matrix created for each GO annotation

The resulting enriched terms were presented using a Word Cloud that can be found here. We can see that among the most prominent molecular functions (in red) are the peroxiredoxin, thioredoxin peroxidase and antioxidant activities, which are functions related to the structure of the protein family and superfamily as noted before. Meanwhile, for the biological processes, defense response to tumor cells and respiratory burst involved in defense response highlight the family's effectiveness against ROSs.

Motifs:

The first step in identifying conserved short motifs was to parse the files found at **ELM classes** and **ProSite patterns** in order to extract motif regular expressions, which were then combined in a single list. Then, the disordered regions for SwissProt were parsed and stored in a python dictionary. Searching for motifs inside of a family was performed by matching the motif regular expressions against every protein sequence, restricted only to intervals representing disordered regions for that sequence. Additionally, all regular expression matches were counted and the counts were stored in a dictionary. Finally, the mean and the standard deviation were computed for the regular expression counts distribution and the

Z-score was calculated for each entry. The motifs with Z-score above 2 were considered highly conserved.

Regular Expression	Count	Z-score
[GSTADNEKR].	164	2.96
[SGADNIT].	130	2.15
[LIVMFYSTAGPC].	141	2.41
[GSTACKRNQ].	151	2.65
[GSTPCEQ].	140	2.39

Table 5. Conserved Motifs with Z-score above 2