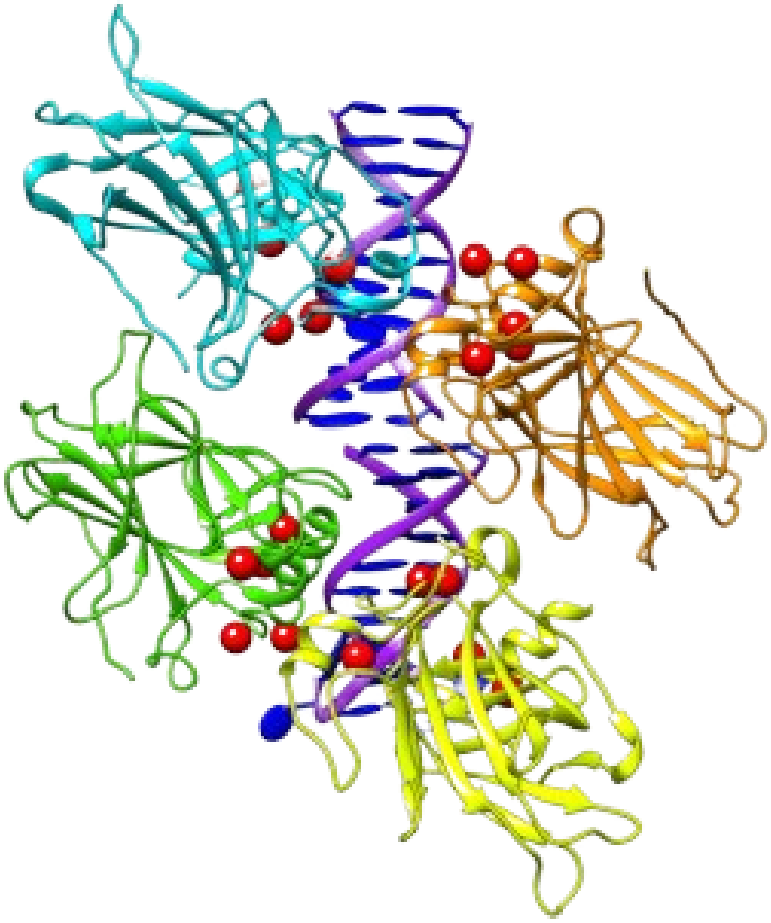


FAMILY MODEL - DOMAIN CHARACTERIZATION

Elisa Ferrero, elisa.ferrero@studenti.unipd.it ; ID: 2089385
Nour Al Housseini, nour.alhousseini@studenti.unipd.it ; ID: 2081230
Hazeezat Adebimpe Adebayo, hazeezatadebimpe.adebayo@studenti.unipd.it ; ID: 2090254

Professor:
Silvio Tosatto
Damiano Piovesan

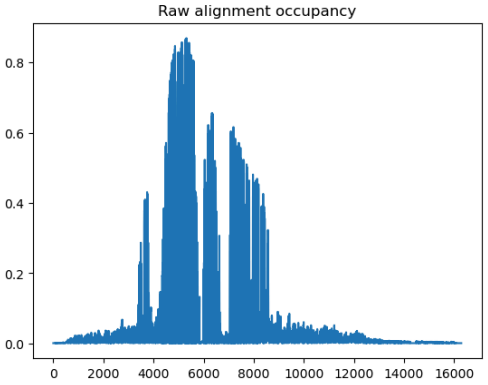
12/26/2024

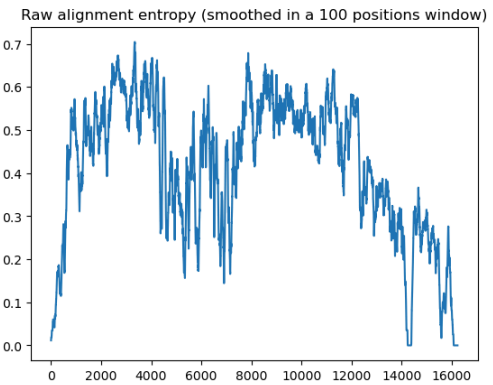
1. Introduction

As architectural engineers, proteins stand proudly in the intricate world of molecular biology. At the heart of protein exploration lies profound characteristics of domains, motifs, and protein families, symbolizing the complex relationships they have in the cellular universe. Domains, akin to distinct functional modules, play a pivotal role in those multifaceted approaches. A quality worth investigating.
In this report, our goal is to decipher the function and characteristics of the protein family defined by the Pfam domain PF00362, with the integrin beta subunit's VWA domain. This module is a protein responsible for recognizing and binding to specific ligands, especially linking the extracellular matrix to the cytoskeleton inside the cell. Not only that, but it plays in various physiological and pathological processes including wound healing, cell signaling, and tissue development. Starting from a Blast search of the domain sequence against Uniref50, we'll define a PSSM and HMM of the domain. We'll then compare their performance, and define a protein family as the proteins found to contain the domain by the best model, in Swissprot. We'll then look at this family, in particular at its taxonomy, GO annotations and functional enrichment, and presence of sequence motifs known as ELM classes and ProSite patterns.

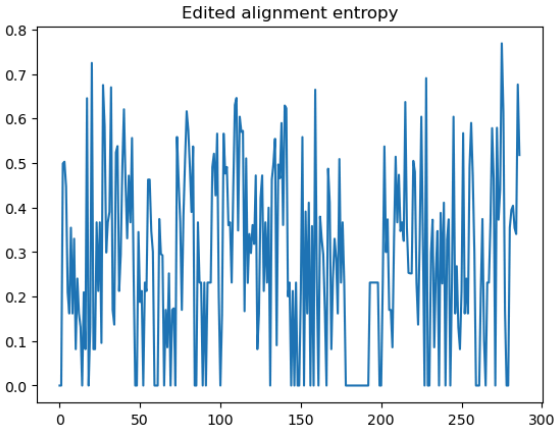
2. Model Building

The model starts with the output of a BLAST search using our assigned seed sequence against the UniRef50 database. UniRef50 is a database that clusters sequences at 50% identity and provides a representative sequence for each cluster. By grouping the protein sequences with more than 50% identity, this process reduces redundancy in the dataset. The search resulted with 1000 reported sequences saved in an xml file. We checked the last match and found it to still be of good quality, with an E-value of 0.00002. We then decided to keep all matches.
Along with other statistics and relevant information, knowledge will be stored in a designated data frame for later analysis. This is followed by a parsing of the accessions lists provided by the Uniref50 search, which will undergo preprocessing such as duplication removal resulting in a cleaned file. The newly processed file will be used as input to retrieve the Uniref50 sequences. To generate a multiple sequence alignment, ClustalOmega was chosen as the proper program due to its effective progressive alignment approach resulting in a hierarchical generation of the input sequences. The output of alignment will be subjected to further processing on Jalview, a designated visualization software characterized by a user-friendly interface.The raw alignment is still incomplete, as it will be subjected to processing steps even further.





By observing the two plots, we can see that the most informative region is clustered around the 4000-6000 mark characterized with low entropy, visualized on the entropy scale. An important quality we are looking for is high occupancy. This stems from the fact that high occupancy indicates that a particular position in the sequence is highly preserved, thus it is of elevated functional importance. Not only that, but the lower entropy will symbolize lower variability among these specific residues. To emphasize the conservation analysis of following sequences, we look for a stretch of a range of 200-300 conserved positions in these regions eventually trimming from the left and right of it. Another important step in processing data. Rows with redundancy of less than or equal to 95% will also be removed, limiting down according to the positions of the domains. We have successfully decreased the sequences from 700 to around 250 sequences, seen in the following edited graph:

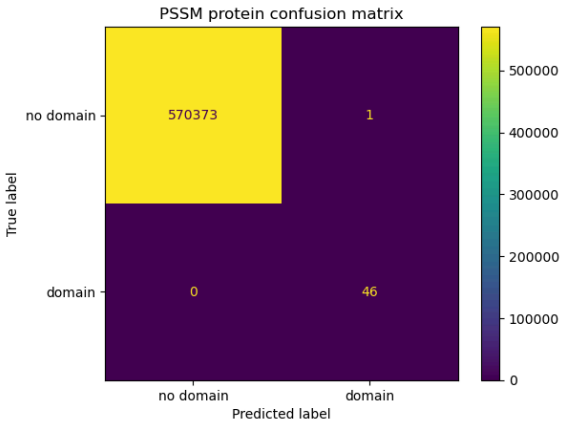


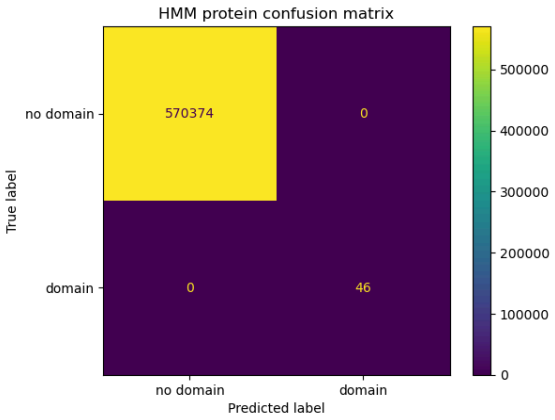
The edited alignments of the MSA provided will be used to generate PSSM and HMM models, which will be discussed in the following segment.

3. Model Evaluation

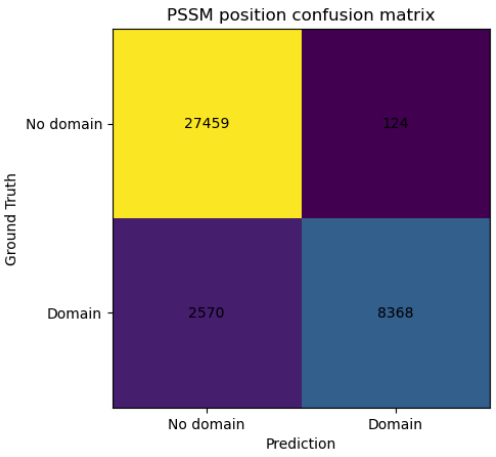
Being important computational models in bioinformatics and computational biology, HMM and PSSM models stand out in their analytical significance. The pipeline starts with collecting data from Swiss-Prot that include all entries with our assigned Pfam domain and having them stored in a designated data frame, representing the ground truth. We performed a PSI-BLAST search against Swissprot and parsed the results to extract all relevant information suitable for our needs. We also performed an hmmscan search using the HMM. The outputted entries will be used against our previous retrieved results to generate predictions about our model's validity. The next step involves building the protein confusion matrices.

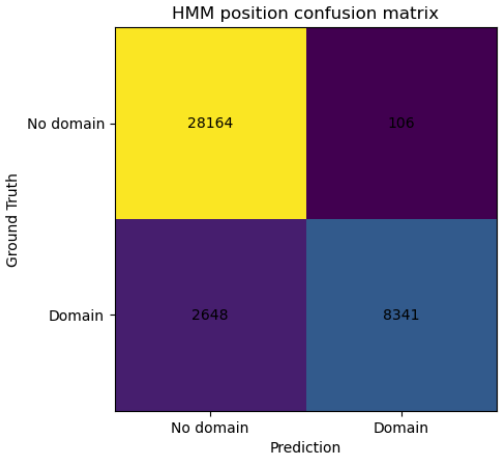
PSSM results reveal that the model was able to correctly predict 46 proteins of true positive nature alongside a single false positive. Meanwhile, the HMM had similar results as it was able to predict 46 true positives. This emphasizes the model's ability in successfully capturing and predicting the proteins that truly possess our domain of interest from an abundant dataset.





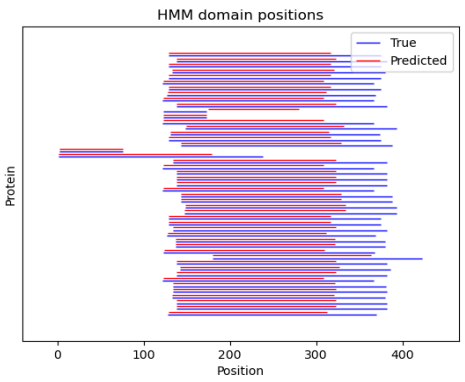
We can conclude that the models act similarly in terms of protein level, comparing at the residue level is next.





As the PSSM predicted 8368 true positives, HMM predicted 106 false positives. The models compensate and go hand in hand in terms of prediction. Not only that, but in terms of performance metrics, the F-score, accuracy, and recall ability are nearly identical with a few decimals different.
Keeping in mind that the HMM was able to avoid false positives at the protein level, sets them apart and thus puts the HMM at a considerable advantage. Thus in the following steps, HMM hits will be used.

As a quick recap for the model evaluation, we decided to visualize predicted HMM hits against their true and ground truth domains as a way to understand how our model analyzed the data.

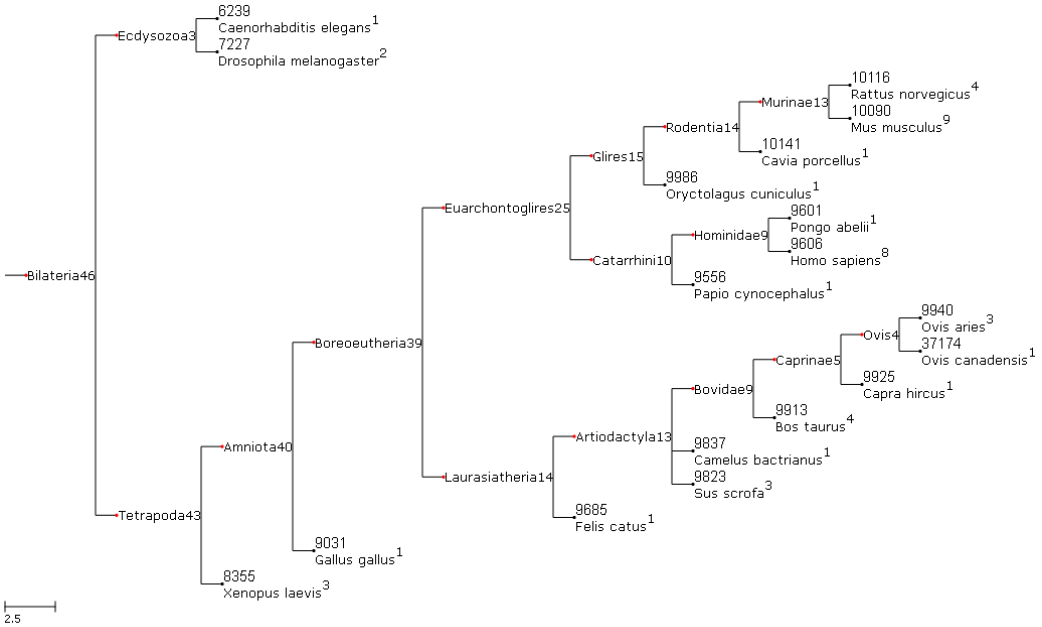


As we can see, the HMM is very good at predicting the start of the domain, but tends to predict an earlier end to it.

Domain Family Characterization

4. Taxonomy

In the taxonomy analysis of our project, we delve into the evolutionary context of the identified protein family. By collecting the taxonomic lineage for each protein within the family from the UniProt XML, we gain insight into the distribution of the domain across different branches of the tree of life. Plotting the taxonomic tree provides a visual representation of the family's diversity, with the number of proteins in a taxon annotated to the right of it. We plotted only the taxa where the tree splits to avoid clutter. This exploration not only enhances our understanding of the domain's evolution but also offers valuable information about its prevalence across various organisms.

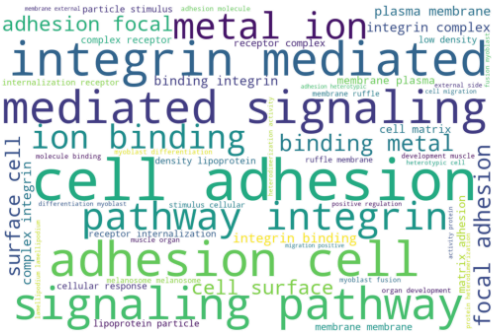


In our model, Bilateria appears as a major node/group of organisms with bilateral symmetry. The deeper we move down the cladogram tree the more diverse the organisms become. We can note that Mus Musculus and Homo sapiens are the taxa with the highest abundance of proteins from our family, with 9 and 8 representatives respectively.

5. Function

The following segment of the model aims to analyze and perform gene ontology enrichment analysis or GO involving several steps. First, we downloaded a tsv file with GO annotations for every swissprot protein from UniProt. Second, enrichment analysis for each term is performed by using Fisher's exact test to determine if there's a significant association between the family and GO term (compared to the GO annotations available in the SwissProt database).

We were able to verify that both two-wailed and right-tail P-values are close to zero, indicating significant enrichment. Not only that, but p-values were adjusted for multiple testing under the false discovery rate control.
The top terms in terms of enrichment are ‘integrin-mediated signaling pathway’, ‘integrin complex’, ‘focal adhesion’ and ‘integrin binding’.
Third, plotting of the enriched terms was performed in a word cloud generated for visualization as seen below.



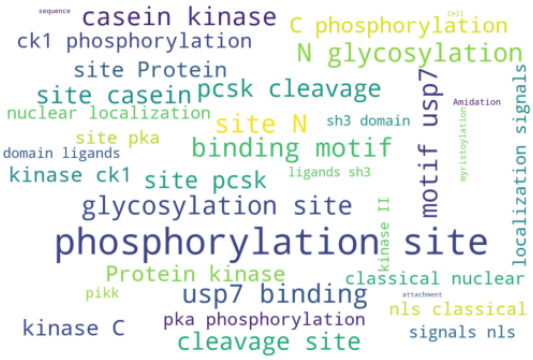
By looking at the word cloud generated,we can note several important functionalities. The term “cell” captures our attention as it suggests a strong association with working inside the cellular environment, whether internally or externally. We see the term ‘integrin’ is prominent, which is expected, since the presence of our domain is part of what defines integrins. We also see that our domain plays a role in signaling pathways involved in cell adhesion, and in the binding of metal ions. This confirms the information found in the Pfam entry of our domain.
We then created a dataset of the top level enriched terms. We started by creating a set of enriched terms that aren't children of other enriched terms, then sorted them by ontology then topologically, thus creating a newly formed dataframe of terms. We have 5 enriched branches.
After topological sorting, the top level terms are:

- “cell adhesion molecule binding” for molecular function;
- “positive regulation of cell migration” for biological process;
- “membrane” for cellular components.

6. Motifs

Motif analysis in disordered regions is a main identifier for comprehensive understanding of protein functionalities in protein families; as they are associated with cell flexibility and various cellular processes.. The model starts with reading data and extracting disordered regions for each family ID from the UniProt database. We also downloaded and parsed the databases of elm classes and prosite patterns. Following up with the analysis, we focus on analyzing the frequency of ELM classes within the disordered regions of the family by looking for matches of regular expressions of elm classes and prosite patterns in the disordered regions. A Z-score for individual classes is calculated based on the frequency distribution of matches, and the matches with a score exceeding 3 will be saved to a separate CSV file (for ELM classes). Processing prosite patterns, and the consecutive

matching to disordered regions are next. Identifying matches between Prosite patterns and disordered regions helps in understanding the presence of specific functional motifs in the family. This information is valuable for functional annotation and characterization of proteins.We opted to visualize the prosite patterns in a word format to provide a sense of the frequency of protein names and descriptions from the Prosite and ELM databases. This will assist in identifying common patterns/ motifs in the datasets as well as their distributions.



“Phosphorylation” and “glycosylation” suggest protein modulation, “cleavage” and “binding” are fundamental for its function, “kinase” and “casein” suggest regulatory events, and finally “localization” is essential for the integrin’s role for cell adhesion and signaling. These features are prominent in the protein family, and are emphasized even more in this case.

7. Results and Conclusion

In this project, we modeled the Pfam domain PF00362 starting only from a seed sequence. Our models demonstrated robust performance in capturing the essential features of the assigned protein domain. The taxonomic analysis unveiled the evolutionary distribution of the protein family, with the taxonomic tree providing a clear representation of its prevalence across diverse organisms. We then analyzed the functions of proteins in our family, finding results coherent with current knowledge of the domain’s function. In particular, we verified its role in activating signaling pathways that promote cell adhesion. We continued to assess the conserved motifs of our family’s disordered regions, finding evidence for conserved phosphorylation and glycosylation sites, and PCSK cleavage sites. This was confirmed by the prosite pattern matches, who also show conserved phosphorylation / glycosylation sites.