Biological Data project

Riccardo Carangelo
riccardo.carangelo@studenti.unipd.it

Fabiana Rapicavoli
fabiana.rapicavoli@studenti.unipd.it

Stefano Minto
stefano.minto.1@studenti.unipd.it

Alina Skrylnik
alina.skrylnik@studenti.unipd.it

February 14, 2023

1 Introduction

"Protein domains are conserved parts on protein sequences and structures, which can evolve, function, and exist independently of the rest of the protein chain. [1]". Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist of several structural domains. One domain may appear in a variety of different proteins.

This project is about the prediction and characterization of a domain family starting from a single sequence. In particular, the protein's sequence we take into consideration is Pyridoxamine kinase/Phosphomethylpyrimidine (Uniprot code: A0A0J9X285) from *Acinetobacter baumannii* (strain IS-123) [2], which covers only a portion of the full sequence of the protein (163 out of 177 amino acids). Pyridoxamine kinase/Phosphomethylpyrimidine belongs to the ribokinase superfamily and it is part of the thiamine pyrophosphate (TPP) synthesis pathway. TPP is an essential cofactor for many enzymes.

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).

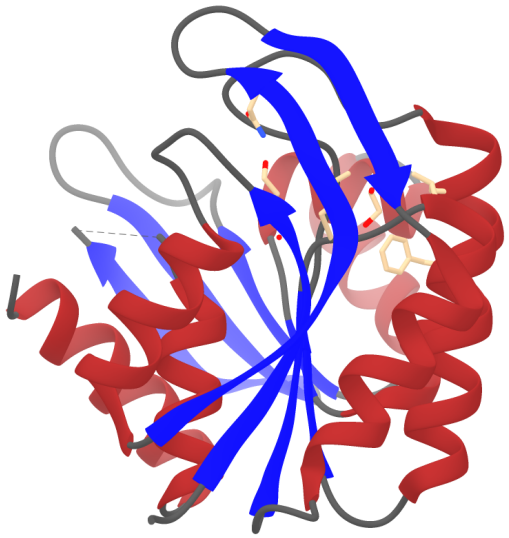

Figure 1: 3D rendering of the A0A0J9X285 protein, showing nine distinct β-strands forming a coiling supersecondary structures partially enveloped by α-helices (rendered using UCSF Chimera [5]).

2 Models Building

We began with a BLASTp search of our sequence against three different databases: Swiss-Prot, UniRef50 and UniRef90. UniRef50 turned out to be the best database, providing 1000 significant hits (maximum visible output in the EBI website) with a high identity values and low E-value scores.

After that, we extracted the accession ID of every hit and retrieved its sequence. The retrieved IDs have been used to find the sequences of the 1000 proteins, found with BLAST, using the mapping UniProt service. Finally, the provided sequences have been used into Clustal Omega to get the MSA (multiple sequence alignment) in FASTA format.

Such MSA has been edited, using the Jalview software, by deleting columns with poor conservation and redundant sequences, setting a 98% tolerance, since if we used a different value (below 90%), almost all sequences would be deleted, since they showed a very high redundance.

3 Models Evaluation

We started with a search using an HMM search against Swiss-Prot database, using the code "PF08543" as Pfam domain. The output, then, was saved (annotated and not-annotated) in a .tsv file and it was was set 0.01 as a threshold for p-value, in order to select 648 meaningful entries.

3.1 MSA editing

To improve the performance of the models, some refinements were necessary. In particular, in order to increase the quality of the alignment, the thresholds of occupancy and entropy were established.

We started from the multiple sequence alignment, using AlignIO module of Biopython, and moved on to calculation of occupancy and entropy. Occupancy is defined as a ratio of amino acids without gaps to the size of the column. The Shannon entropy is calculated with the scipy module, taking as input the probabilities of amino acid appearance.

The resulting data frame was filtered by a threshold value of 0.5 for occupancy, which means that at least one half of the hits are actually aligned, and a value of 0.3 for entropy.

This edited MSA was used to create HMM and PSSM models, using local software (BLAST and HMM). From the position column in increasing order, we established the range, used then in Jalview, in order to produce an improved MSA. In particular, we eliminated the portion of MSA, which is in the left of first position and the portion that is in the right of the last position, and the MSA file remained with a length of 4171 (last position) - 2883 (first position) = 1288 aligned positions.

After all these operations, the number of sequences were reduced from 1000 to 39.

We then eliminated from that MSA edited file the redundant sequence and low conserved columns. Thanks to these modifications, we reached a MSA which showed a good level of performance.

3.2 Comparison at the protein level

We carried out the HMM search, using the HMM model created before. Regarding the PSSM model, by using PSI BLAST and reiterating one time the search, until reaching a number of results as close as possible to the threshold value (number of hits obtained by using Pfam domain ID against Swiss-Prot database, which are 648), 625 hits were provided with only one iteration.

Using these two files, it was possible to make a comparison of the selected hits obtained from HMM search against Swiss-Prot search and the hits obtained from HMM and PSSM models search.

The results obtained from the two methods are displayed in the following table:

	True Positives	False Positives	False Negatives
HMM	406	13	242
PSSM	486	136	147

Note: the number of true negatives are not relevant, as they include all other Swiss-Prot entities.

In order to compare the results of the two models, the following accuracy metrics were exploited:

$$Specificity\ (TNR) = \frac{TN}{TN + FP}$$

Specificity (True Negative Rate): it refers to the proportion of true negative cases, which are correctly identified by a test or a model.

$$Precision\ (PPV) = \frac{TP}{TP + FP}$$

Precision (Positive Predicted Value): it is a measure of the accuracy of a diagnostic test or a predictive model. It represents the proportion of positive cases, which are correctly identified as such by the test or the model.

$$Recall\ (TPR) = \frac{TP}{TP + FN}$$

Recall (True Positive Rate or sensitivity): it refers to the proportion of true positive cases, which are correctly identified by a test or a model.

$$F1{-}score = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$$

F1-score: it is the harmonic mean of precision and recall and it provides a single value, which summarizes the accuracy of the test or the model. The F1-score is a useful metric for evaluating the performance of a test or a model, when there is an imbalance between the number of positive and negative cases, as it gives equal weight to precision and recall.

$$BA = \frac{TPR + TNR}{2}$$

Balanced accuracy: it represents a modification of the traditional accuracy metric, which gives equal weight to both positive and negative cases, regardless of their frequency.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Matthews Correlation Coefficient: it is a measure of the accuracy, which takes into account true and false positives and negatives. It is a balanced metric, which considers both positive and negative predictions of a model and provides a single value, that summarizes its accuracy.

The results obtained with respect to the entire Swiss-Prot database having 568.744 entries are:

	HMM	PSSM
Specificity score:	1	0.999
Precision score:	0.969	0.782
Recall / sensitivity score:	0.627	0.769
F1-score:	0.761	0.776
Balanced accuracy:	0.813	0.884
Matthews Correlation Coefficient:	0.779	0.775

From this table, it is possible to notice that specificity, precision scores and Matthews Correlation Coefficient are higher when using HMM search, whereas PSI-BLAST performed better in terms of recall, F1-score and balanced accuracy.

The biggest difference is shown by precision (18.66%), recall (-14.24%) and balanced accuracy (-7.1%). We can conclude that the HMM search provided more relevant elements, since the number of false positives was very low, while PSI-BLAST showed higher ratio of actual true values.

3.3 Comparison at the residue level

In a similar way as in the subsection 3.2, we compared the results with the selected hits from Swiss-Prot, after the HMM and PSI-BLAST search.

A comparison between the starting/ending target alignment positions between HMM and Pfam and between PSI-BLAST and Pfam was made, using Pfam search as a ground truth. In particular, we checked if starting position of HMM/PSI-BLAST hit is greater or equal to the starting position of Pfam hit and the ending position of HMM/PSI-BLAST hit is lower or equal to the Pfam hit.

The accuracy metrics of both models are provided below:

	HMM	PSSM
Specificity score:	0.999	0.999
Precision score:	0.871	0.517
Recall / sensitivity score:	0.576	0.508
F1-score:	0.693	0.512
Balanced accuracy:	0.788	0.754
Matthews Correlation Coefficient:	0.708	0.512

In the residue level, all of the metrics are higher when exploiting HMM search. The biggest difference is shown by the precision score (35.46%), F1-score (18.09%), and Matthews Correlation Coefficient (19.61%).

4 Domain family and characterization

4.1 Taxonomy

From the family sequences obtained by PSI-BLAST search using the PSSM, that gave better scores than HMM-search, we extracted the sequences by batches of 100 sequences, and for each protein a taxonomic lineage from the Proteins API EMBL-EBI was extracted [3]. Then, a dictionary was built, with the protein IDs as keys and the lists of taxonomic lineages as values.

From the ID of each protein, it was possible to obtain the taxonomic tree, using the Bio package Phylo and Networkx.

Four different types of trees have been plotted, one single dendrogram-like taxonomic tree showing the evolutionary relationships among the species related to the proteins of the model and three other trees for representing separately Bacteria, Eukaryota, and Archaea taxa with a Kamada-Kawai layout and with nodes of sizes proportional to the frequency of each taxon. Due to their large sizes, all trees can be seen in full resolution in the supplementary material.

Taxonomic results showed that most of the proteins provided by PSSM-based model were bacterial proteins, mostly found in *Proteobacteria* and *Firmicutes* phyla. Despite this prevalence, it was also possible to find some hits belonging to *Archaea* and *Eukaryota* organisms. The latter taxon showed mostly *Fungi* proteins, even if some human and also model organisms' hits appeared, like *Mus musculus* and *Danio rerio*.

4.2 Function

An OBO file was parsed with a specific function which returned a list of ontologies. From that file, we obtained a dictionary with GO terms as keys and the lists of unique ancestors as values. By performing this step, three roots were found, corresponding to the different namespaces: "Molecular function", "Biological process" and "Cellular component". Then we uploaded a .csv file with all Swiss-Prot entries with the GO terms associated.

From this file, it was possible to obtain unique GO terms for each protein and to implement a function, called propagation, which allowed to check the correspondence between unique GO terms for each protein and the dictionary with the ancestors of each GO. Reiterating this process for each GO term of each protein allowed to create a new dictionary with the protein IDs as keys and the associated GO terms as values (directly and not).

Using the last dictionary, it was possible to select the proteins which belong to our PSSM-model, obtain all the unique GO terms, associated to our model and compute the confusion matrices and fold increase for each GO term.

The top ten of GO terms with the highest fold increase are:

GO term	Fold increase	Description
GO:0008972	7841.74	phosphomethylpyrimidine kinase activity
GO:0008902	5690.91	hydroxymethylpyrimidine kinase activity
GO:0033785	4852.86	heptose 7-phosphate kinase activity
GO:0052855	4784.84	ADP-dependent NAD(P)H-hydrate dehydratase activity
GO:0033786	4692.01	heptose-1-phosphate adenylyltransferase activity
GO:0008478	4498.27	pyridoxal kinase activity
GO:0006796	3906.58	phosphate-containing compound metabolic process
GO:0006793	3797.12	phosphorus metabolic process
GO:0004747	3557.58	ribokinase activity
GO:0009443	3441.74	pyridoxal 5'-phosphate salvage

We then computed the Fisher exact test for each confusion matrix, in order to select only the relevant terms, eliminating the ones with a p-value lower of equal than 0.05 for one-sided test and lower than 0.025 for two-sided test.

In the following plot, it is possible to see the enriched terms in a word cloud (Figure 2), with the words size proportional to the fold increase value, in order to obtain a visualization of the selected GO terms.

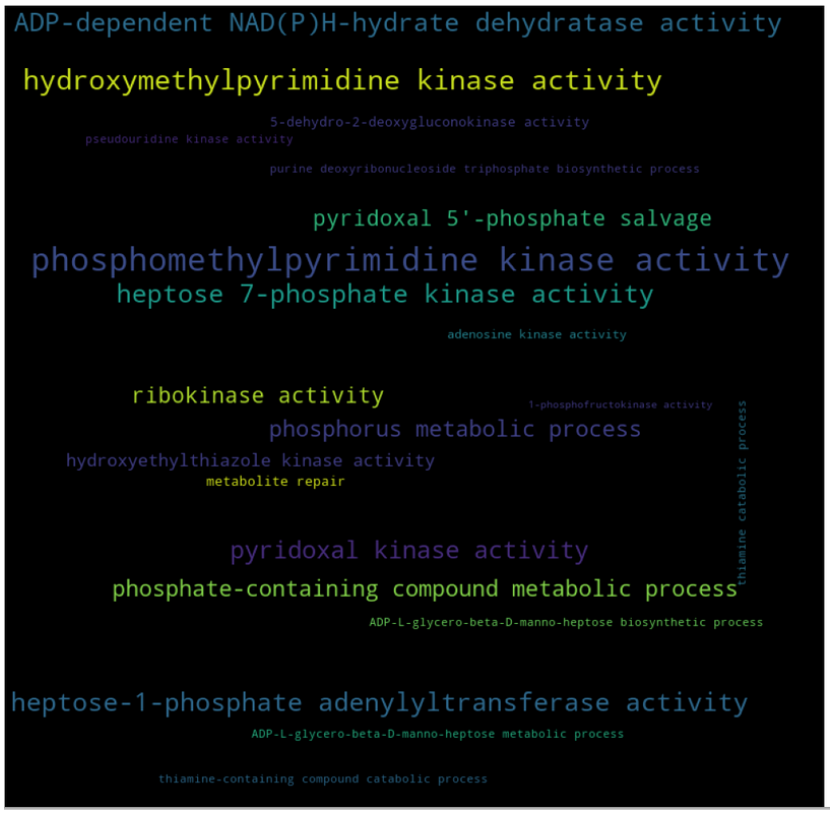

Figure 2: Word cloud plot highlighting most common functions in the selected model, by showing most enriched terms.

It can be highlighted by these results that the main activity of this domain, i.e., pyrimidine-like

molecules kinase, even if it is possible to see also a high enrichment for low p-value terms, representing transferase activities and dehydratase activities.

In order to obtain the enrichment of branches, we gathered all the direct children terms for each GO term, by exploiting the parsed OBO file. Using the depth first search algorithm, it was possible to obtain all the possible branches in our ontology. Finally, we looped over each branch to look for enriched terms and computed the mean value of fold increase for each branch (Figure 3).

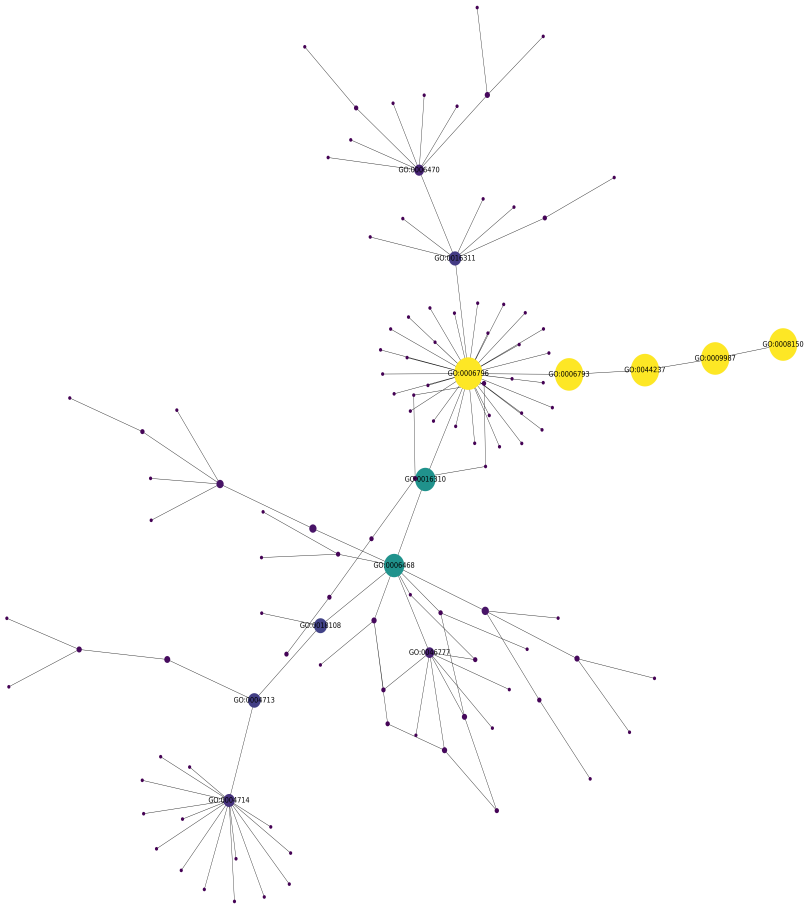

Figure 3: Graph showing most enriched branched and the more representative enriched branches (biggest node sizes).

4.3 Motifs

In this part, we looked for significantly conserved short motifs inside the disordered region in our family, included in the PSSM-based prediction model.

In order to do that, some specific lines was exploited from ProSite patterns. ProSite is a database of protein family domains which consists of regular patterns, representing significant biological sites. These sites are defined along the protein sequence contiguously and through a regular grammar.

We also used ELM classes [4], where ELM stands for Eukaryotic Linear Motif. This computational biology resource mainly focuses on annotation and detection of eukaryotic linear motifs, by providing both a repository of annotated motif data and an exploratory tool for motif prediction. ELMs are

compact protein interaction sites composed by short stretches of adjacent amino acids. They are enriched in intrinsically disordered regions of the proteome and provide a wide range of functionality to proteins.

We loaded a .tsv file coming from Swiss Prot, containing all the labeled proteins with disordered regions and their position. Then, this uploaded file has been used to filter the disordered regions in the sequences of our PSSM-based model. Every disordered region was compared with the known ProSite and ELM's patterns and all the different pattern matches found were counted.

These are the most frequent motifs from Prosite:

Accession	Functional site name	Frequency
PS00005	Protein kinase C phosphorylation site	2
PS00006	Casein kinase II phosphorylation site	2
PS00004	cAMP- and cGMP-dependent protein kinase phosphorylation site	1
PS00008	N-myristoylation site	1

These are the ten most frequent motifs from ELM:

Accession	Functional site name	Frequency
ELME000012	di Arginine retention/retrieving signal	2
ELME000443	Polo-like kinase phosphosites	2
ELME000417	14-3-3 binding phosphopeptide motif	2
ELME000288	IAP-binding motif (IBM)	1
ELME000528	Adaptin binding Endosome-Lysosome-Basolateral sorting signals	1
ELME000135	WW domain ligands	1
ELME000365	WDR5 WD40 repeat (blade 5,6)-binding ligand	1
ELME000501	CIN85 and CD2AP SH3 domain binding motif	1
ELME000155	SH3 ligand	1
ELME000006	SH3 ligand	1

From the tables we can notice that all frequencies are low in both cases. This makes sense, considering we have found just 8 distinct proteins with small disordered regions during the filtering step.

Anyway, the patterns found are coherent with the predicted domain function and taxonomy, since several kinase (or, more in general, phosphorous) activity processes can be observed. Finally, ProSite domains, if compared to ELM domains, showed a functional site name closer to the actual function found during the functional analysis.

This is probably due to the fact that our starting sequence is Prokaryotic and that, also, most of the hits in the final PSSM-based model are bacterial proteins. Therefore, it is plausible to expect less interesting results while looking to ELM, which contains only Eukaryotic linear motifs.

5 Conclusion

Even if functional definition and characterization of a protein is a challenging field, it is easy to notice what a powerful analysis method can be offered by bioinformatics tools, which supply a complete and sharp set of instruments to perform the task.

Using several databases and tools, it is possible to provide an accurate enough model for a specific protein sequence, using it to look up homologous proteins and to predict and characterize a protein family for input sequences. The overall function of the model, the taxonomic distribution, and other characteristics can be inferred using an *in silico* bioinformatics-driven method, with surprising results that can constitute a powerful way lay the foundations for orienting further laboratory studies.

References

1. National Library of Medicine, National Center for Biotechnology Information
2. InterPro, a protein database

3. Proteins API EMBL-EBI
4. Eukaryotic Linear Motif resource for Functional Sites in Proteins
5. UCSF Chimera - UCSF Resource for Biocomputing, Visualization, and Informatics
6. HMMER - Builds HMM models of multiple sequence alignments. Performs HMM/sequence database searches
7. BLAST is an alignment software used in that specific case for build the PSSM
8. JalView - Multiple sequence alignment viewer
9. Clustal-Omega - Multiple sequence alignment
10. NCBI-BLAST - Perform database sequence searches
11. UniProt - Protein Database