Modelling and Functional Characterization of the Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase Domain Family

Marco Uderzo
Department of Mathematics, University of Padua
marco.uderzo@studenti.unipd.it
ID: 2096998

Tanner Graves
Department of Mathematics, University of Padua
tanneraaron.graves@studenti.unipd.it
ID: 2073559

Claudio Palmeri
Department of Mathematics, University of Padua
claudio.palmeri@studenti.unipd.it
ID: 2062671

Abstract

This project aims to build a sequence model and provide a comprehensive functional characterization of the Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase domain family. The models' accuracy is benchmarked against Pfam annotations in the SwissProt database. Furthermore, we delved into the functional and structural properties of the domain family, analyzing the taxonomic lineage, assessing Gene Ontology (GO) annotations for functional enrichment, and searching for significantly conserved short motifs inside the family. In particular, our investigation into conserved short motifs using ELM and ProSite databases identified several significant motifs, such as the WDR5 WD40 Repeat-binding ligand and the N-myristoylation site.

1 Introduction

1.1 Protein Domains

A *protein domain* represents a conserved part of a protein's sequence and three-dimensional structure, capable of evolving, functioning, and existing independently from the rest of the protein chain. These domains, each forming a stable and compact 3-D structure, are essential components in proteins, often occurring in various combinations across different proteins. Domains are fundamental in molecular evolution, serving as versatile building blocks that can be rearranged to form proteins with diverse functions. This adaptability and independence make them crucial in understanding protein structure and function.

1.2 Pyridoxamine Kinase / Phosphomethylpyrimidine Kinase

The *Pyridoxamine Kinase / Phosphomethylpyrimidine Kinase*[1] family is a group of enzymes that play key roles in various biochemical pathways, particularly in the metabolism of vitamins and coenzymes. This family includes two distinct but related enzymes:

- *Pyridoxamine Kinase*: This enzyme is involved in the vitamin B6 metabolism pathway. Vitamin B6 exists in different forms, including pyridoxamine, pyridoxal, and pyridoxine. Pyridoxamine kinase specifically catalyzes the phosphorylation of pyridoxamine, converting it into pyridoxamine 5'-phosphate. This is an important step in the salvage pathway of vitamin B6, which is crucial for its recycling and maintenance within the cell.

- *Phosphomethylpyrimidine Kinase*: This enzyme plays a role in the biosynthesis of thiamine (vitamin B1), which is essential for numerous cellular functions, particularly in carbohydrate metabolism.

Both these enzymes, due to their roles in vitamin metabolism, are crucial for maintaining cellular health and function. Disruptions in these pathways can lead to vitamin deficiencies, affecting numerous biological processes.

1.3 Objective of the Study

In this project, our primary objective is to construct and refine a sequence model for the *Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase* domain family, and to characterize its functional aspects. To ensure the reliability and accuracy of our models, we are aligning and comparing them against the established Pfam annotations within the *SwissProt* database. We then delve into the domain family's functional and structural attributes. This includes a detailed analysis of their taxonomic lineage, providing insights into their evolutionary history and biological diversity. Additionally, we are assessing the Gene Ontology (GO) annotations. This process is crucial for identifying functional enrichment within the family and understanding the broader biological roles these domains play. Furthermore, we are focused on detecting and analyzing significantly conserved short motifs. The identification of these motifs is essential as they often play critical roles in the domain's functional properties and interactions within the cell.

2 Domain Model Definition

2.1 Model Building

Firstly, we investigated the target family to model - *Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase* - and verified that the provided representative `A0A0J9X285`[2] protein sequence, having Pfam domain `PF08543`, is indeed characteristic of the protein family. This was done by retrieving the seed alignment used to generate the HMM defining the Pfam family from *InterPro*, and aligning the representative query sequence to the seed alignment using *JalView*.

The query spans the length of the seed alignment and the gaps opened in the query correspond to low occupancy regions in the seed alignment. This bolstered our confidence that performing a homology search with our query sequence would have been able to return sequences belonging to the `PF08543` family. This was done by performing a Position-Specific Iterated BLAST (PSI-BLAST) search on *SwissProt*. The results were downloaded as a `.fasta` file and opened in *JalView*, where we added our query sequence as a reference. The FASTA file was aligned with the query sequence using *Clustal Omega*.

The query sequence overlapped the primary conserved regions of the MSA, and the majority of positions outside of the query had very low occupancy, consisting of sequences that were unusually long. The query bounds for the MSA are observed to be reasonable bounds to trim the MSA, so positions outside this range were trimmed from it.

Sequences that opened gaps more than a couple residues long were investigated by referencing the BLAST hit corresponding to that sequence. Many of these instances were from Eukaryotes - which is atypical for this family - and were of reasonable quality. Since it is useful to include this information, no sequences reported by BLAST were discarded.

The MSA was finalized by removing the query sequence, and it was then processed to generate a *Position-Specific Scoring Matrix (PSSM)* using the command line `PSI-BLAST` tool, with the *SwissProt* database as the reference. Finally, the *Hidden Markov Model (HMM)* was built using the `hmmer hmmbuild` command.

2.2 Model Evaluation

The PSSM predictions were generated through PSI-BLAST searches against the SwissProt database. Parallelly, we used hmmer hmmsearch, and the results were parsed to extract alignments between the HMM and sequences in the SwissProt database. It is worth noting that HMM search initially returned far more results than sequences in the family, so we chose an `E-value` threshold of 10^{-38} that minimizes the false positive proteins identified by the HMM model to good effect.

Both of the models predict regions in sequences corresponding to our family. We evaluated their performance on two levels:

- Protein-level: are sequences containing the family correctly identified;
- Residue-level: are the positions within these sequences correctly identified.

2.2.1 PSSM Protein-Level Performance Evaluation

The protein-level performances of the PSSM model are shown in the table below:

Metric	Value
Precision	0.973
Recall	0.48
F1-Score	0.643
Balanced Accuracy	0.740
MCC	0.683

Table 2.1: PSSM Evaluation at Protein-Level

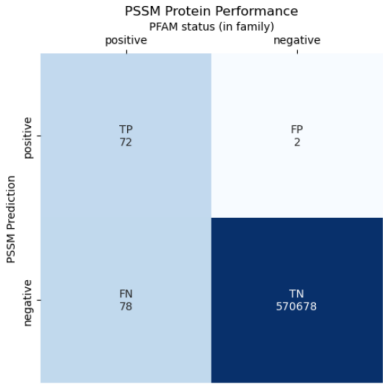


Fig. 2.1: Confusion Matrix for PSSM at protein-level

2.2.2 PSSM Residue-Level Performance Evaluation

The residue-level performances of the PSSM model are shown in the table below:

Metric	Value
Precision	0.980
Recall	0.491
F1-Score	0.655
Balanced Accuracy	0.746
MCC	0.694

Table 2.2: PSSM Evaluation at Residue-Level

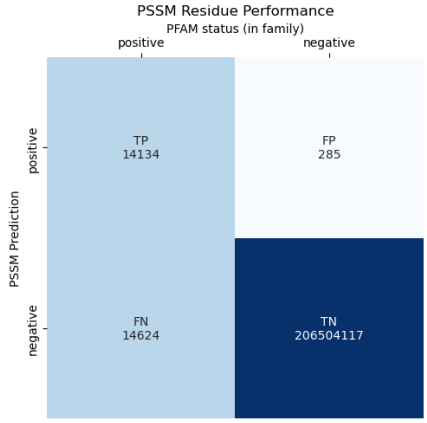


Fig. 2.2: Confusion Matrix for PSSM at residue-level

2.2.3 HMM Protein-Level Performance Evaluation

The protein-level performances of the HMM model are shown in the table below:

Metric	Value
Precision	0.993
Recall	0.993
F1-Score	0.993
Balanced Accuracy	0.996
MCC	0.993

Table 2.3: HMM Evaluation at Protein-Level

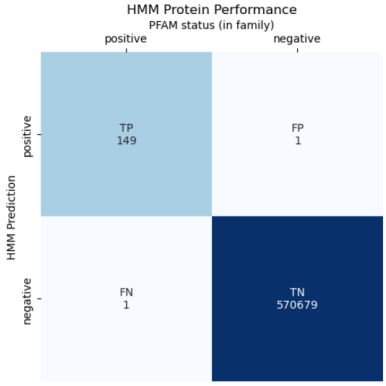


Fig. 2.3: Confusion Matrix for HMM at protein-level

2.2.4 HMM Residue-Level Performance Evaluation

The residue-level performances of the HMM model are shown in the table below:

Metric	Value
Precision	0.989
Recall	0.974
F1-Score	0.982
Balanced Accuracy	0.987
MCC	0.982

Table 2.4: HMM Evaluation at Residue-Level

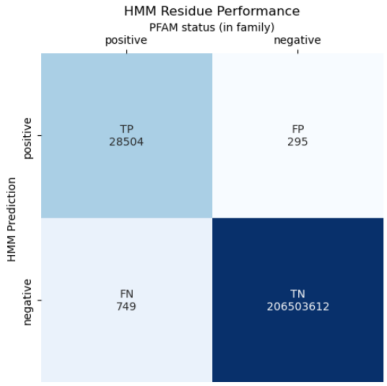


Fig. 2.4: Confusion Matrix for HMM at residue-level

Both models score very high precision which is due to the fact that the dataset is highly skewed toward negative matches. Only the HMM model is capable of also obtaining a high recall value - and overall much better statistics - thus matching the target PFAM family very closely while the PSSM model only recognizes about half of the target family. Therefore, only the HMM model was used for the next steps of the project.

3 Domain Family Characterization

3.1 Taxonomy

To construct the taxonomic tree, we assembled the lineage data derived from the *SwissProt* database, corresponding to the proteins identified by our HMM model as belonging to our family. The lineages were used to generate a comprehensive taxonomic tree, which was enriched with node-specific information, including taxonomic names and the frequency of each taxon's occurrence within our data. In our tree, the size of each node indicates how many examples (or leaves) have that taxonomy term. This provides a good visualization of the lineage of taxonomy terms characteristic of our family (i.e. Bacteria, Pseudomonadota, Gammaproteobacteria, Enterobacterales, Enterobacteriaceae, E. coli).

The taxonomic tree is shown below. In order to view it in full resolution, we refer to the corresponding section 5.1 in the appendix, or directly to the `TaxonomyTree.pdf` file that can be found in the supplementary material.

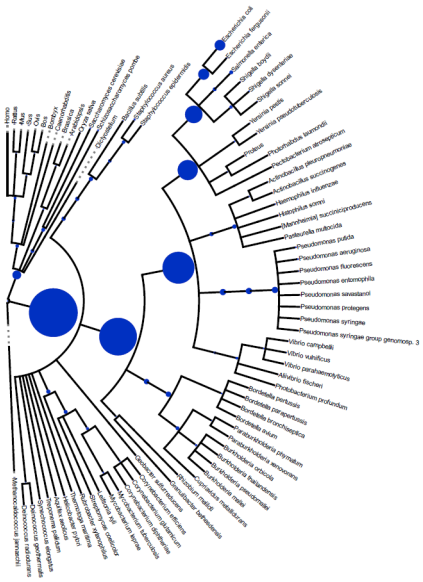


Fig. 3.1: Taxonomic Tree

3.2 Functional Enrichment with Gene Ontology Annotation

We performed Functional Enrichment Analysis using *Gene Ontology* (GO) annotations by extracting the *molecular function*, *cellular component*, and *biological process* data.

In order to visualize which GO Terms are characteristic of our family, we can plot the enrichment (probability) of observing a GO Term over both our model family and the totality of *SwissProt*. Selecting the terms with the highest odds, or the ratio of probability that the term is observed in the family and all of *SwissProt* gives us clues about which terms are most characteristic.

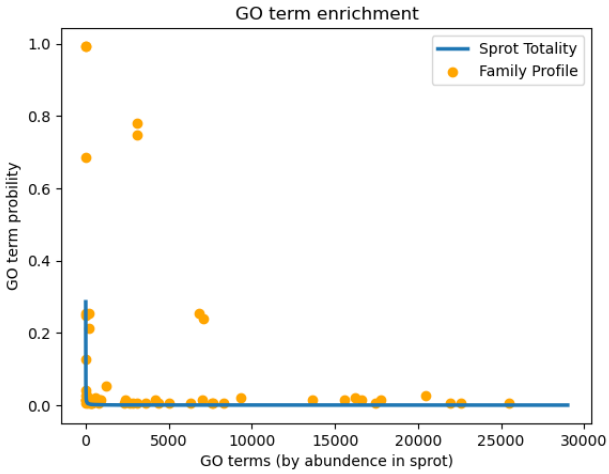


Fig. 3.2: Enrichment of GO Terms

GO Term ID	Term Name	Odds
GO:0042817	pyridoxal metabolic process	3805.53
GO:0008478	pyridoxal kinase activity	3773.28
GO:0009443	pyridoxal 5'-phosphate salvage	3642.90
GO:0008972	phosphomethylpyrim. kin. act.	3615.25
GO:0008902	hydroxymethylpyrim. kin. act.	3605.24
GO:0009230	thiamine catabolic process	1902.76
GO:0042818	pyridoxamine metabolic process	1427.07
GO:0042816	vitamin B6 metabolic process	1268.51
GO:0010054	trichoblast differentiation	1268.51
GO:0036172	thiamine salvage	1087.29
GO:0042822	pyridoxal phosphate metab. proc.	951.38
GO:0070280	pyridoxal binding	845.67
GO:0031403	lithium ion binding	634.25
GO:0042819	vitamin B6 biosynthetic proc.	543.64
GO:0050334	thiaminase activity	456.66
GO:0097159	organic cyclic compound binding	200.29
GO:0008614	pyridoxine metabolic process	131.22

Table 3.1: Most characteristic GO Terms

3.2.1 Fisher's Exact Test

The Fisher-exact test is useful for testing the nonrandom association between two categorical variables. In our case, possessing a GO term and membership into our family of proteins. In order to conduct the test contingency table for a single GO term of interest must be constructed:

# in family with term	# in family without term
# outside fam. with term	# outside fam. without term

The test reporting a low p-value is evidence of the studied GO term being strongly associated with membership in our family. And indeed, we observe that for the terms with the highest odds, the p-value is extremely close to 0.

Below, we plot a word cloud of the Enriched Terms for each aspect:

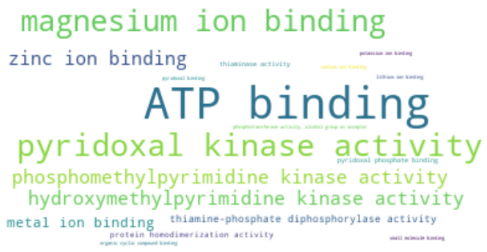


Fig 3.3: Enriched terms for aspect: Molecular Function

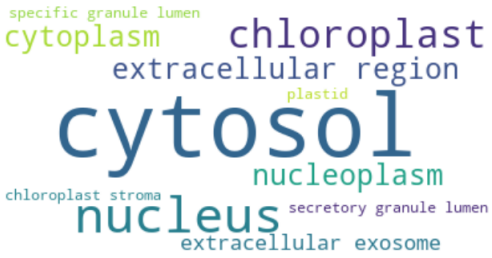


Fig 3.4: Enriched terms for aspect: Cellular Component

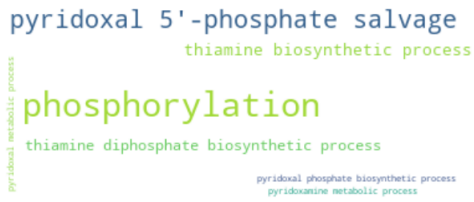


Fig 3.5: Enriched terms for aspect: Biological Process

We then explored the hierarchical structure of the Gene Ontology to discern the most significantly enriched branches relevant to our protein family.

High-level terms in Gene Ontology, typically positioned at the top of the hierarchy, encapsulate the broadest categories, covering a diverse range of specific functions, processes, or components. These general terms are less detailed compared to low-level terms, but are instrumental in offering an overarching view of the primary biological functions, processes, or components linked with a set of genes or proteins.

Leveraging the GO hierarchy, each GO Term was classified according to its level of specificity. We focused on high-level terms, filtering the GO terms based on their hierarchical level. This approach enabled us to identify the most enriched branches at a more generalized level, revealing key biological processes, molecular functions, and cellular components prominently involved in our protein family. We report them in the table below:

GO Term ID	Term Name	Dom.	Prob.
GO:0005829	cytosol	cc	0.253
GO:0005576	extracell. region	cc	0.013
GO:0005654	nucleoplasm	cc	0.013
GO:0097159	org. cyc. comp. binding	mf	0.013
GO:0036094	small molecule binding	mf	0.013

Table 3.2: Most enriched branches
cc: cellular component; mf: molecular function

3.3 Motifs

Motifs are short protein sequences that are often repeated across the genome. These motifs usually coordinates protein-to-protein interaction and are found in the disordered regions.

Our objective is to see if any commonly occurring linear motifs appear in our `PF08543` protein family and to do so we have at our disposal two datasets: ELM and ProSite.

For each member of the family we checked if the regular expressions found in the aforementioned two datasets were sub-sequences of it. To do that, we used a precompiled file[3] which correlates *ProSite* entries and their corresponding patterns translated to regular expressions. However, the vast majority of the matches found are outside of disordered regions, which can be found with the *MobiDB-lite* database. This is due to the fact that our protein family is constituted by globular proteins.

Given a motif, the regions where its pattern matches our proteins are overlayed onto Multiple Sequence Alignments. These patterns are then visually inspected to determine the significance of pattern in the family. Some patterns are overly general, matching many regions and are

labeled as having a high probability of being observed in any given protein sequence from our family.

Conservation of a pattern in the same position is indicative of functional significance.

3.3.1 ELM

There were 18 significant hits in the *ELM* Database. Below, we present the most common one.

WDR5 WD40 Repeat (blade 5,6)-binding Ligand

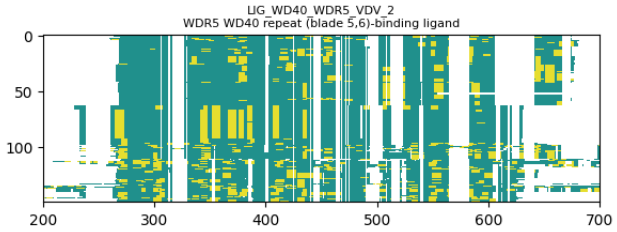


Fig 3.7: Most common Elm motif in our family

The WD40 repeat is a short structural motif, approximately 40 amino acids in length, often terminating in a tryptophan-aspartic acid (W-D) dipeptide. Proteins that contain WD40 repeats are known as WD-repeat proteins. These repeats are typically involved in protein-protein interactions and can form a beta-propeller structure. Each repeat forms a blade of the propeller, with typically 4 to 8 repeats coming together to form a circular propeller-like structure. WDR5 is a protein containing WD40 repeats, which form a beta-propeller structure, crucial for protein-protein interactions. It plays a pivotal role in the regulation of gene transcription, cell cycle progression, and DNA repair through its involvement in post-translational modifications of histones. Specifically, WDR5 plays a key role in H3K4 methylation and H4K16 acetylation by acting as a scaffold protein for the assembly of the respective core histone methylation and acetylation complex. The WDR5-binding motif, as described by the ELM database [4] (accession `ELME000365`), is a functional site that interacts with WDR5 between blades 5 and 6 of the WD40 repeat. The WDR5-binding motif is conserved across various species, indicating its importance in biological processes.

Plots for the other 17 motifs can be found in the appendix section 5.2.

Other common motifs found are:

- *MOD_GlcNHglycan*: A Glycosaminoglycan attachment site. Glycosaminoglycan are long, linear polysaccharides consisting of repeating disaccharide units. These can be found attached via a serine residue to proteoglycans which are extracellular proteins found at the cell surface and in the extracellular matrix. The glycosaminoglycan attachment site is an exposed serine which accepts transfer of xylose. [5]
- *LIG_FHA_1*: A FHA phosphopeptide ligand. FHA are small domains that form a sandwich of two antiparallel beta sheets. The FHA domain is a signal transduction module which recognizes phosphothreonine containing peptides on the ligand proteins and has a role in: cell-cycle checkpoint control, DNA repair, signal transduction, transcriptional regulation, and pre-mRNA splicing.[6]
- *CLV_PCSK_SKI1_1*: A PCSK cleavage site. The subtilisin-like proprotein convertases play a major role in the proteolytic processing of both neuropeptide and peptide hormone precursors. PCSK1 (proprotein convertase 1) and PCSK2 (proprotein convertase 2) are type I proinsulin-processing enzymes that play a key role in regulating insulin biosynthesis.[7]

It is also worth noting that the *ELM* patterns have identified Casein kinase 1 & 2 Phosphorylation sites, however they do not appear to be as well preserved than as reported by the ProSite patterns. This can likely be attributed to the ELM patterns being too specific and not accommodating sequence variation in the site, but more investigation is needed.

Also interestingly, many patterns are preserved but may appear in different regions. This is one advantage of plotting the patterns on MSAs, as trying to detect the preservation of these patterns at the position level would miss this.

3.3.2 ProSite

There were 4 significant hits in the *ProSite* Database. We report the most common one below:

N-myristoylation Site

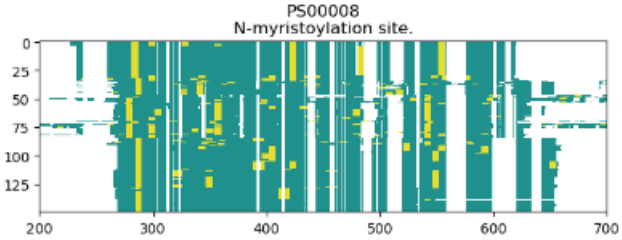


Fig 3.8: Most common ProSite motif in our family

N-myristoylation consists in the acylation of eukaryotic proteins by the covalent addition to their N-terminal

residue of myristate, a C14-saturated fatty acid [8]. N-myristoylation has been observed happening in animals, plants, fungi, protozoans and viruses. It allows for weak protein–protein and protein–lipid interactions and plays an essential role in membrane targeting, protein–protein interactions and functions widely in a variety of signal transduction pathways, for example in:

- *Apoptosis*: The myristoylation of pro-apoptotic BH3-interacting domain death agonist (Bid) leads to the production of cytochrome c which causes cell death.
- *Cancer*: Increased myristoylation of c-Src gene can lead to enhanced cell proliferation and be responsible for transforming normal cells into cancer cells

The other 3 significant hits in the *ProSite* database are the following:

- *PS00001*: A N-glycosylation site where oligosaccharides are attached to eukaryotic proteins by binding themselves to the N-terminal asparagine residue.[9]
- *PS00005*: A protein kinase C phosphorylation site, where protein kinase C manage the phosphorylation of serine or threonine residues near the C-terminal residue.[10]
- *PS00006* A casein kinase II phosphorylation site, where Casein kinase II (CK-2) is a protein serine/threonine kinase whose activity is independent of cyclic nucleotides and calcium. CK-2 functions as a regulator of signal transduction pathways.[11]

Plots for the top 4 most common motifs in *ProSite* can be found in the appendix at section 5.3.

4 Conclusion

In this project, we have successfully constructed a sequence model for the Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase domain family and provided a detailed functional characterization. Our model, benchmarked against Pfam annotations in the SwissProt database, demonstrated high accuracy in identifying the domain sequences. We analyzed the taxonomic lineage of the domain family, which provided insights into their evolutionary history and diversity. The assessment of Gene Ontology annotations revealed the functional enrichment within the family, highlighting their biological roles and processes.

Additionally, our investigation into conserved short motifs using ELM and ProSite databases was a key aspect of the project. We identified several significant motifs, such as the WDR5 WD40 Repeat-binding ligand and the N-myristoylation site.

References

[1] InterPro Database. Pyridoxamine kinase/phosphomethylpyrimidine kinase. https://www.ebi.ac.uk/interpro/entry/InterPro/IPR013749/.

[2] UniProt Database. Phosphomethylpyrimidine kinase. https://www.uniprot.org/uniprotkb/A0A0J9X285/entry.

[3] GitHub Stevin Wilson. Prositepatternstopython-regex. https://github.com/stevin-wilson/PrositePatternsToPythonRegex/tree/master.

[4] ELM Database. Wdr5 wd40 repeat (blade 5,6)-binding ligand. http://elm.eu.org/elms/LIG_WD40_WDR5_VDV_2.html.

[5] ELM Database. Mod_glcnhglycan. http://elm.eu.org/elms/MOD_GlcNHglycan.html.

[6] ELM Database. Lig_fha_1. http://elm.eu.org/elms/LIG_FHA_1.html.

[7] ELM Database. Clv_pc5k_5ki1_1. http://elm.eu.org/elms/CLV_PCSK_SKI1_1.html.

[8] ProSite Database. N-myristoylation site. https://prosite.expasy.org/PDOC00008.

[9] ProSite Database. N-glycosylation site. https://prosite.expasy.org/PDOC00001.

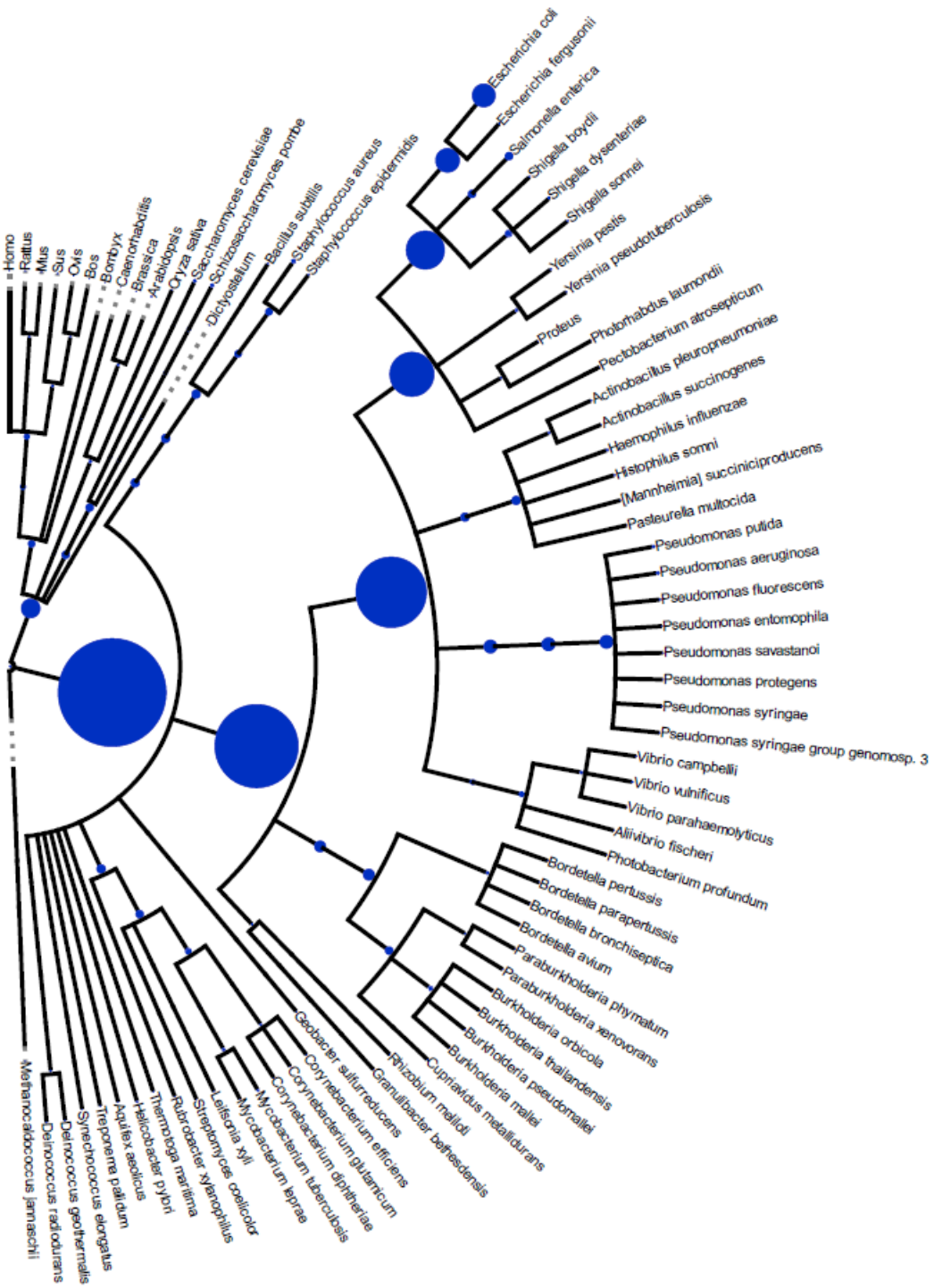
[10] ProSite Database. Protein kinase c phosphorylation site. https://prosite.expasy.org/PDOC00005.

[11] ProSite Database. Casein kinase ii phosphorylation site. https://prosite.expasy.org/PDOC00006.

5 Appendix

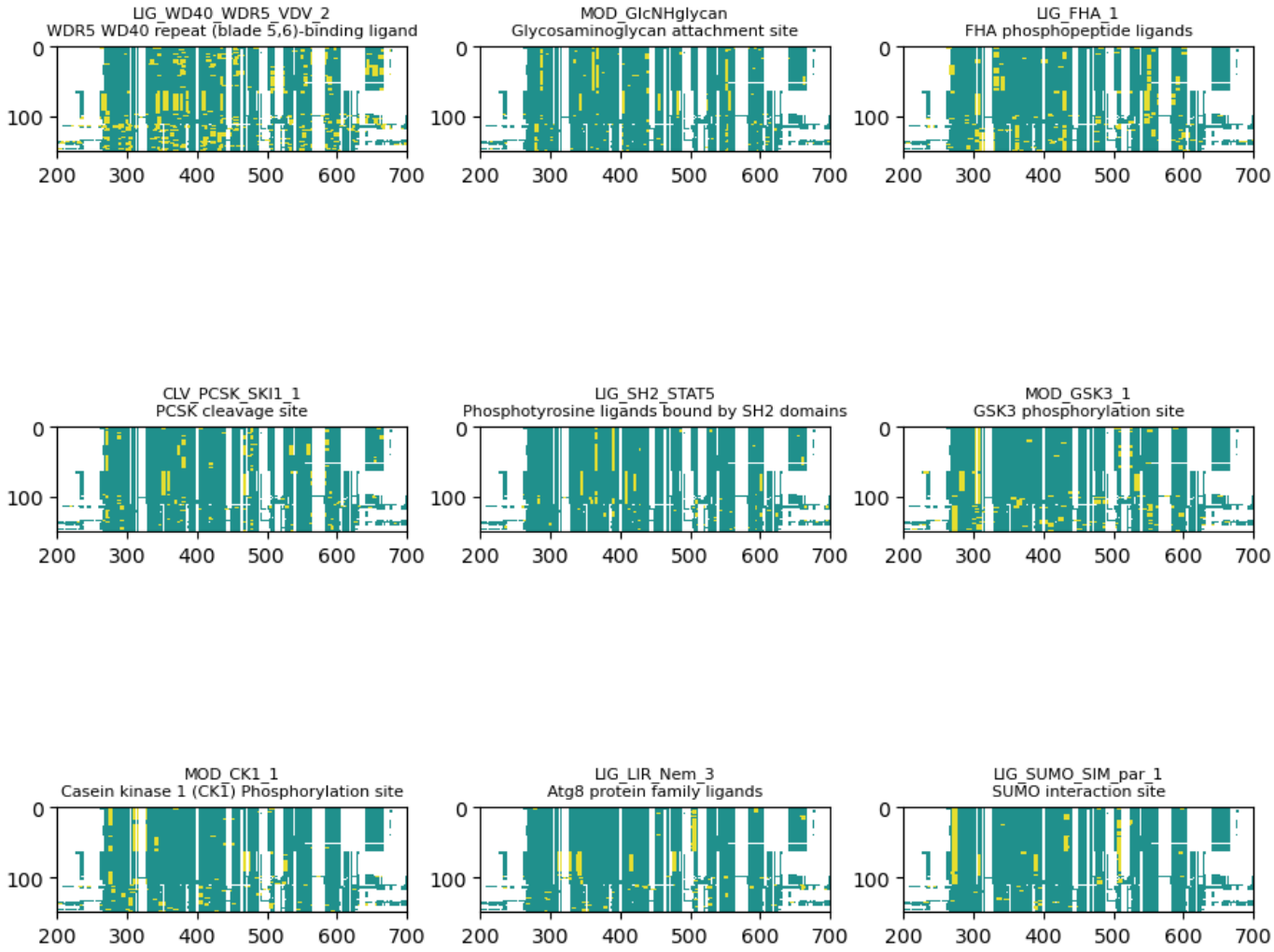
5.1 Taxonomic Tree

Below, we plot the full-resolution Taxonomic Tree.

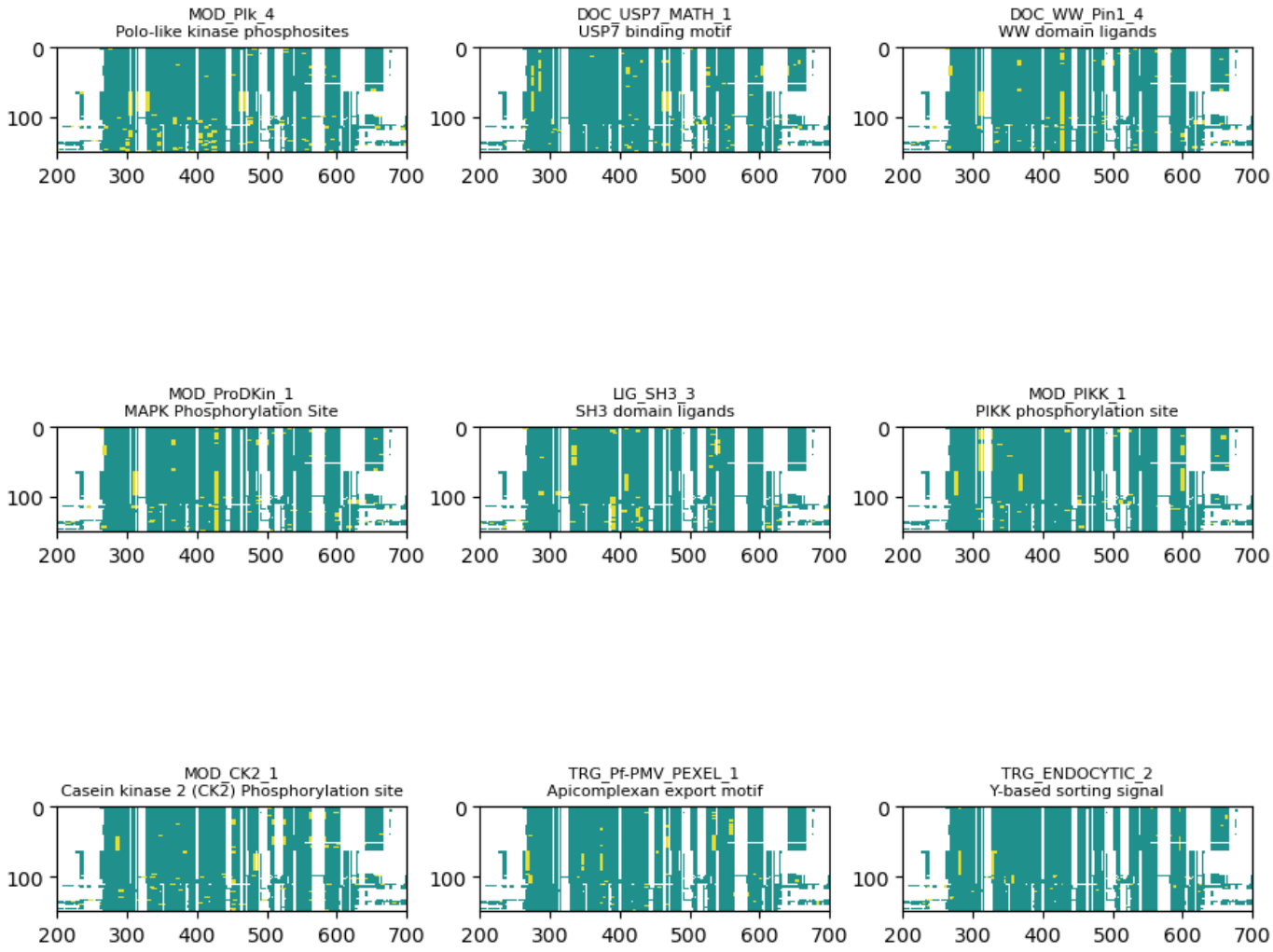


5.2 ELM

Below, we plot the top 9 most common ELM Linear Motifs.



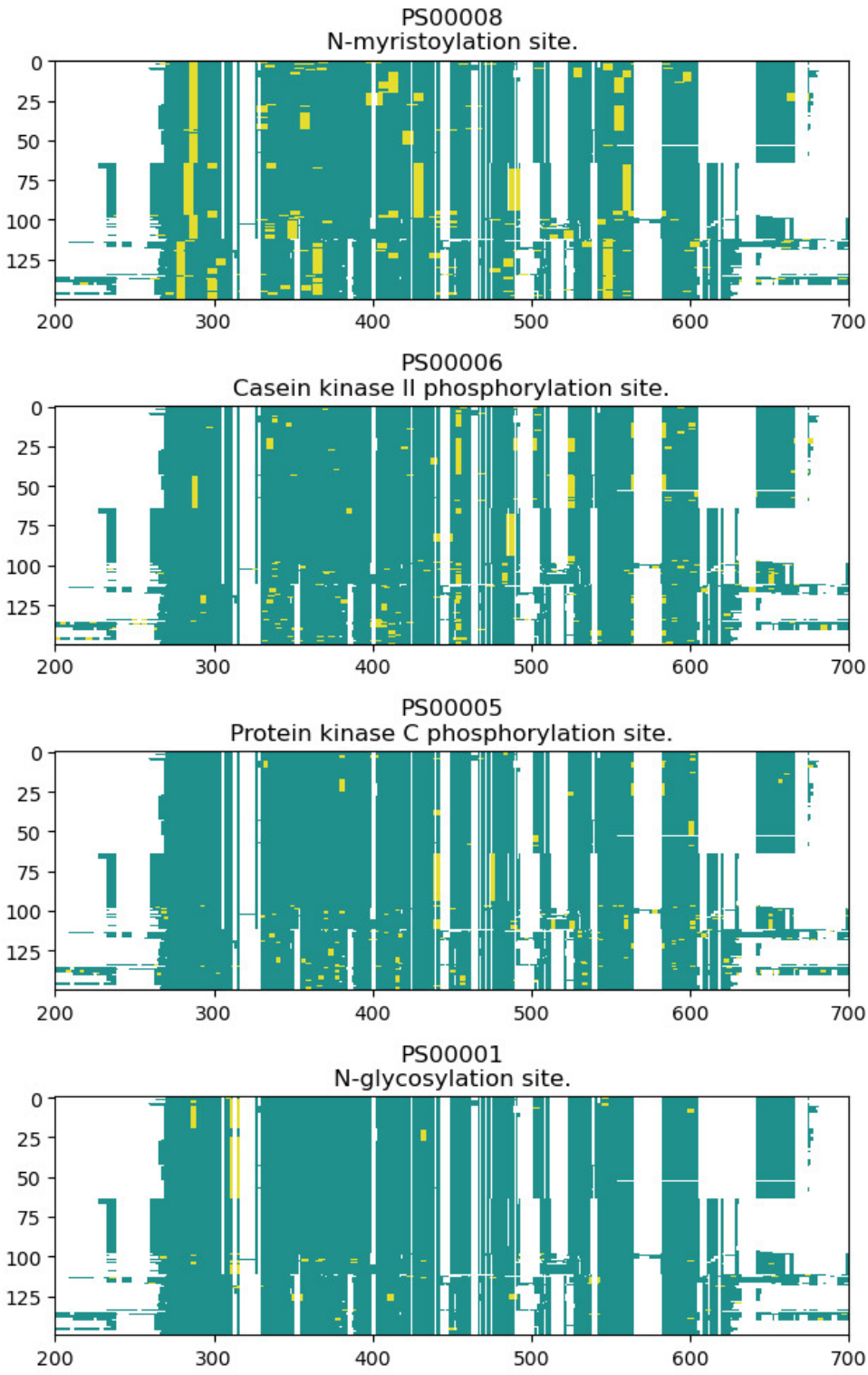
Top 9 most common ELM linear motifs



10th to 18th most common ELM linear motifs

5.3 ProSite

Below, we plot the top 4 most common ProSite Linear Motifs.



Most common 4 ProSite linear motifs