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, this project delves into the functional and structural properties of the domain family, involving taxonomic lineage analysis, Gene Ontology (GO) annotations assessment, and motif searching. (Include findings in the abstract)

1 Introduction

1.1 Protein Domains

In molecular biology, 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 three-dimensional 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(CHECK FOR ERRORS)

I'd need to check if that's factually correct. Check PdxK and ThiD

Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase 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 is a part of the thiamine (vitamin B1) biosynthetic pathway. It catalyzes the phosphorylation of hydroxymethylpyrimidine (HMP) to hydroxymethylpyrimidine phosphate. This step is essential in the synthesis of thiamine pyrophosphate (TPP), an active form of vitamin B1. TPP is a vital coenzyme in several enzymatic reactions, particularly those involved 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

This project aims to build a sequence model and provide a comprehensive functional characterization of the Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase domain family. (Write a small preamble of the goals of this project, even if it is similar to what is written in the abstract).

2 Methods and Results

2.1 Model Building

Firstly, we investigated the target family to model - Pyridoxamine Kinase/Phosphomethylpyrimidine Kinase - and verified that the provided representative `A0A0J9X285` 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 bolsters our confidence that performing a homology search with our query sequence will be 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 sequences 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 Eukariotes - 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 HMM was build using the `hmmer hmmbuild` command.

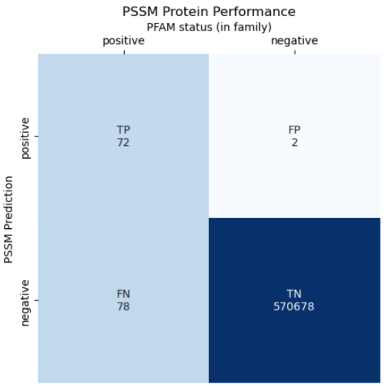
2.2 Model Evaluation

The PSSM predictions were generated through PSI-BLAST searches against the SwissProt database. Parallelly, HMM searches were conducted, the results of which were parsed to extract alignments between the HMM and sequences in the SwissProt database.

2.2.1 PSSM Protein-Level Performance Evaluation

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

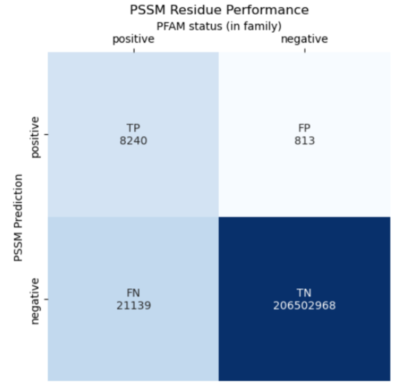
Metric	Value
Precision	0.894
Recall	0.227
F1-Score	0.361
Balanced Accuracy	0.613
MCC	0.45



2.2.2 PSSM Residue-Level Performance Evaluation

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

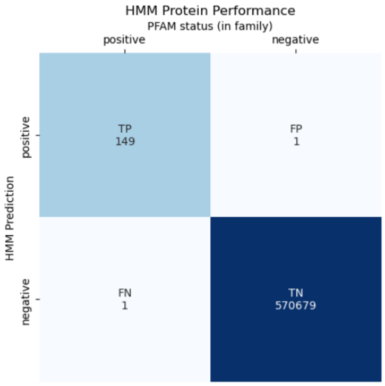
Metric	Value
Precision	0.91
Recall	0.28
F1-Score	0.429
Balanced Accuracy	0.64
MCC	0.505



2.2.3 HMM Protein-Level Performance Evaluation

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

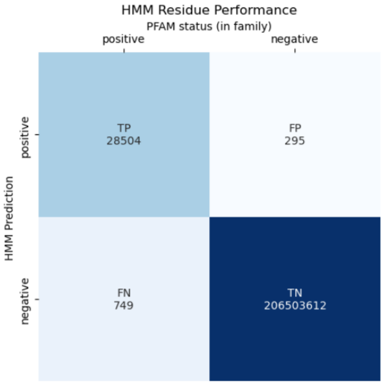
Metric	Value
Precision	0.993
Recall	0.993
F1-Score	0.993
Balanced Accuracy	0.996
MCC	0.993



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

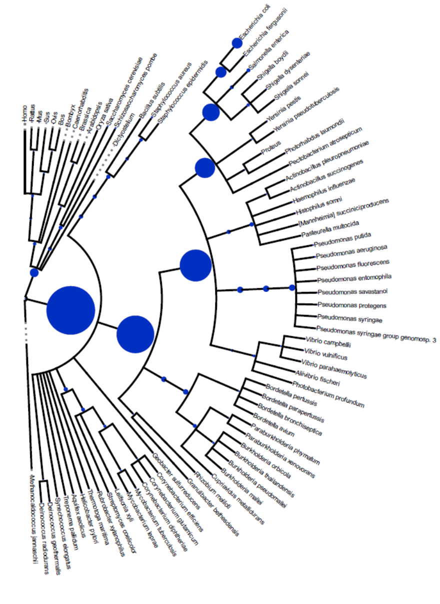


As indicated by the confusion matrix, the HMM predictions match the target PFAM family very closely.

2.3 Taxonomy

To construct the taxonomic tree, we assembled the lineage data derived from the *SwissProt* database, corresponding to the protein family under investigation. The lineages were used to generate a comprehensive taxonomic hierarchy, 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).

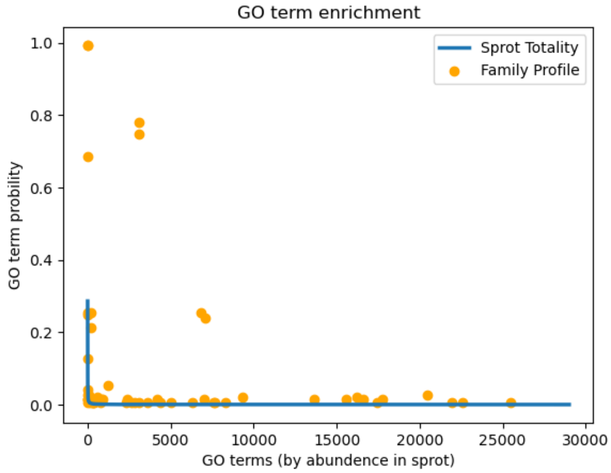
The taxonomic tree is shown below. In order to view it in full resolution, we refer to the corresponding `TaxonomyTree.pdf` file that can be found in the supplementary material.



2.4 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.

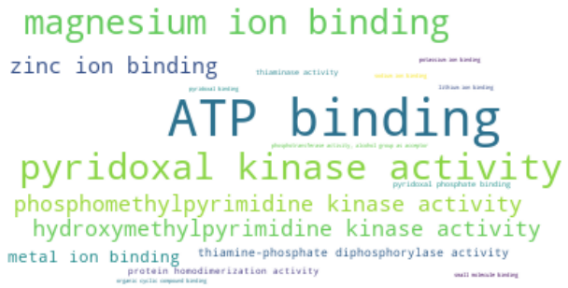


GO Term ID	Term Name	Odds
GO:0042817	pyridoxal metabolic process	3805.53
GO:0008478	pyridoxal kinase activity	3773.28
GO:0009443	pyridoxal 5'-phosphate salvage	3642.90
GO:0008972	phosphomethylpyrim. kin. act.	3615.25
GO:0008902	hydroxymethylpyrim. kin. act.	3605.24
GO:0009230	thiamine catabolic process	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

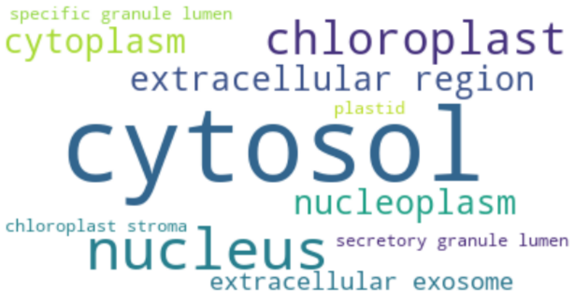
Most characteristic GO Terms

By using Fisher's Exact Test, we realized that the p-value is extremely close to zero (maybe add p=...) for terms with high odds, indicating that they are indeed characteristic of our family. However, as a consequence of how sparse GO Labels are for a sequence, when compared (to the abundance of our limited amout of terms (56) present in our family. ?? -¿ check this).

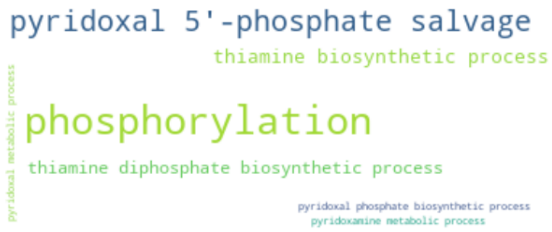
Below, we plot a word cloud of the Enriched Terms for each aspect:


magnesium ion binding
zinc ion binding
thiaminase activity
potassium ion binding
sodium ion binding
pyridoxal binding
lithium ion binding
ATP binding
phosphotransferase activity, alcohol group as acceptor
pyridoxal kinase activity
pyridoxal phosphate binding
phosphomethylpyrimidine kinase activity
hydroxymethylpyrimidine kinase activity
metal ion binding
thiamine-phosphate diphosphorylase activity
protein homodimerization activity
organic cyclic compound binding
small molecule binding

Molecular Function Terms



Cellular Component Terms


pyridoxal 5'-phosphate salvage
thiamine biosynthetic process
pyridoxal metabolic process
phosphorylation
thiamine diphosphate biosynthetic process
pyridoxal phosphate biosynthetic process
pyridoxamine metabolic process

Biological Process Terms

Most enriched branches:

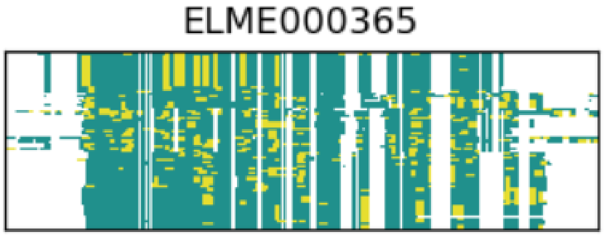
2.5 Motifs

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.

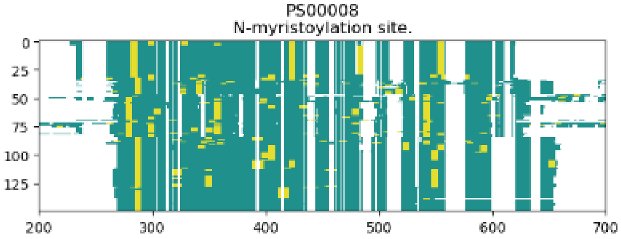
Our objective is to see if any commonly occurring linear motifs appear in our `PF08543` protein family and to do so we have at our disposal 2 datasets: ELM and ProSite.

For each member of the family we checked if the regular expressions found in the aforementioned 2 datasets were sub-sequences of it and counted how many times they appeared. We can use this to find the most common motifs among the entire protein family.

We can then plot the sequences alignment and highlight where a single motif appear to get a visual representation of how frequent it is. This when applied to the most commonly found motifs yields the following:



Most common ELM linear motif



Most common ProSite linear motif

3 Results

4 Discussion

5 References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example. Where appropriate, include the name(s) of editors of referenced books.