Functional and Structural Characterization of a protein domain family

Alberto Calabrese
Data Science, Department of Mathematics, University of Padova
Email: alberto.calabrese.2@studenti.unipd.it

Marlon Helbing
Data Science, Department of Mathematics, University of Padova
Email: marlonjoshua.helbing@studenti.unipd.it

Lorenzo Baietti
Data Science, Department of Mathematics, University of Padova
Email: lorenzo.baietti@studenti.unipd.it

***Abstract*—This study focuses on the characterization of the protein domain family associated with Pfam identifier `PF00151`, specifically the Lipase/vitellogenin domain in *Homo sapiens*. We constructed a Position-Specific Scoring Matrix (PSSM) and a Hidden Markov Model (HMM) to represent the domain, evaluated them against SwissProt annotations, and analyzed their functional and structural aspects. The models effectively captured conserved sequence features, and predictions matched well with known annotations. The results emphasize the domain's critical biological roles and suggest avenues for further refinement and exploration.**

I. Introduction

The domain of a protein is an important part to model, as it can evolve, function and exist independently of the rest of the protein chain [1]. Therefore, in this work, we aim to characterize the domain of a *Homo sapiens* protein that is linked to Lipase/Vitellogenin by building a sequence model starting from a single sequence and providing a functional characterization of the entire domain family. This report is structured as follows. In Section II, we describe how we built both a PSSM [2] and an HMM [3] model, which are then evaluated in their performance on the protein and residue level against SwissProt [4] proteins annotated with the 'ground truth' Pfam domain [5] in Section III. Next, we looked at functional and structural properties of the entire protein family, such as taxonomy in Section IV, function in Section V, and finally motifs in Section VI. In Section VII we report results and a conclusion is provided in Section VIII.

II. Model Building

Given our initial data (Table 1) and the representative domain sequence (Fig. 1), we collected 1000 homologous sequences by means of a BLAST search [6] against the UniProt Knowledgebase (UniProtKB) with an e-value threshold of 0.0001. Next, we utilized ClustalOmega [7] to generate a multiple sequence alignment. To have a more generalizable and performant model later on, we cleaned the MSA by first removing redundant rows at a 100% threshold using JalView [8], resulting in 155 sequences, and then performing a detailed conservation analysis. In particular, we first removed columns of residues that had 80% or more gaps, which resulted in roughly removing 85% of the initial $3,254$ columns. To further prepare the input for the PSSM model, some additional adjustments were required due to excessive gaps at the beginning and end of many sequences, which caused an error during the PSSM generation process. To resolve this issue, we removed columns from the beginning and from the end of the MSA. Based on our observations, removing columns until there are no more than 50 consecutives gaps allowed the PSSM algorithm to function correctly. However, before removing columns, we identified sequences with an unusually high number of gaps compared to the rest. To minimize excessive column removal in the MSA, we first eliminated sequences with more than 79 % of gaps. Given the filtered MSAs, we then proceeded to build both the HMM and PSSM models using NCBI-BLAST-2.16.0+ and HMMER-3.4, respectively.

TABLE 1: Protein Domain Information

Property	Value
UniProt ID	P54315
PfamID	PF00151
Domain Position	18-353
Organism	Homo sapiens (Human)
Pfam Name	Lipase/vitellogenin

Fig. 1: Domain Sequence

KEVCYEDLGCFSDTEPWGGTAIRPLKILPWSPEKIGTRFLLYTN
ENPNNFQILLLSDPSTIEASNFQMDRKTRFIIHGFIDKGDESWVT
DMCKKLFEVEEVNCICVDWKKGSQATYTQAANNVRVVGAQV
AQMLDILLTEYSYPPSKVHLIGHSLGAHVAGEAGSKTPGLSRIT
GLDPVEASFESTPEEVRLDPSDADFVDVIHTDAAPLIPFLGFGTN
QQMGHLDFFPNGGESMPGCKKNALSQIVDLDGIWAGTRDFVA
CNHLRSYKYYLESILNPDGFAAYPCTSYKSFESDKCFPCPDQGC
PQMGHYADKFAGRTSEEQQKFFLNTGEASNF

III. Model Evaluation

To generate predictions, we used PSI-BLAST (1 iteration) and HMM-SEARCH with default parameters. Both searches were performed against the manually curated SwissProt database (release 29.11.2024), resulting in 83 predicted proteins with e-values < 0.001 and the corresponding domain locations within each protein sequence. To generate the ground truth, we collected all 82 reviewed proteins in SwissProt annotated with the given Pfam domain utilizing the InterPro API [9]. We evaluated both models against the ground truth using two approaches. First, at the protein level, we verified whether predicted proteins matched the annotated ones. Second, at the residue level, for proteins present in both our predictions and the SwissProt reviewed set, we created binary vectors representing domain positions and compared them to quantify the overlap of domain boundary predictions. To assess the performance of the model, we calculated precision, recall, F-score, balanced accuracy and MCC. Results are reported in section VII.

IV. Taxonomy

To explore the taxonomic distribution of proteins within our family, we systematically collected and analyzed lineage data for the 82 family sequences. We queried the UniProt API [10] to retrieve the complete taxonomic lineage for each protein of our family, capturing their hierarchical classification from the broadest domain (*Eukaryota*) to the most specific organism. The taxonomic tree was generated using the ETE Toolkit [11] and visualized in the Newick tree format [12]. Node sizes in the phylogenetic tree were adjusted to reflect the relative abundance of each taxonomic entity, with larger nodes indicating higher representation in the protein family. Fig. 2 illustrates the resulting phylogenetic tree, highlighting the distribution of proteins across various taxonomic groups.

V. Function

First, we obtain GO annotations [13] for both our family proteins and the entire SwissProt database. To have a more complete representation, we then expanded these GO annotations by parsing the ontology tree [14] and adding the ancestor GO terms of each initially found GO term. In order to calculate the enrichment of each GO term in our family compared to the ones found in the SwissProt database, we used Fisher's exact test [15]. We used a 2×2 contingency table where rows indicated the presence or absence of a GO term and columns differentiated between proteins within and outside our domain family (Table 2). For each

	Protein in family	Protein not in family
Has GO term	a	b
No GO term	c	d

TABLE 2: Contingency table

GO term, we tested two hypotheses. Under the null hypothesis, the proportion of proteins annotated with that given GO term in our domain family equals the proportion in the full SwissProt dataset. We evaluated this against two alternative hypotheses: a right-tailed test to identify terms that appear more frequently in our family compared to SwissProt, and a two-tailed test to detect any significant differences in proportions in either direction. The enrichment value was then calculated as:

$$\frac{family\_proportion}{swissprot\_proportion}$$

We generated a word cloud visualization of the enriched terms (i.e. the terms, where $p < 0.05$ for both $p_{\text{two-tailed}}$ and $p_{\text{right-tailed}}$ tests) weighted by their enrichment value (Fig. 3). Furthermore, we identified the most significantly enriched branches in the Gene Ontology hierarchy by analyzing enrichment patterns across different levels. Starting from the most specific terms, we moved systematically upward through the ontology, level by level. At each step, we applied the true path rule by assigning each parent term the highest enrichment value found among its immediate child terms. This approach allowed us to trace the most enriched pathways from specific terms to broader biological categories. The 10 most enriched branches can be seen in Table 5.

VI. Motifs

To identify conserved short motifs within our protein family, we began by analyzing the locations of intrinsically disordered regions. These regions were predicted using MobiDB-lite [16]. We then extracted motif patterns from two widely-used resources: ELM classes [17] and ProSite patterns [18]. For ProSite patterns, a script adapted from Stevin Wilson's `PrositePatternsToPythonRegex` [19] project was used to convert motif patterns into Python-compatible regular expressions. This approach allowed us to efficiently identify motifs present in disordered regions.

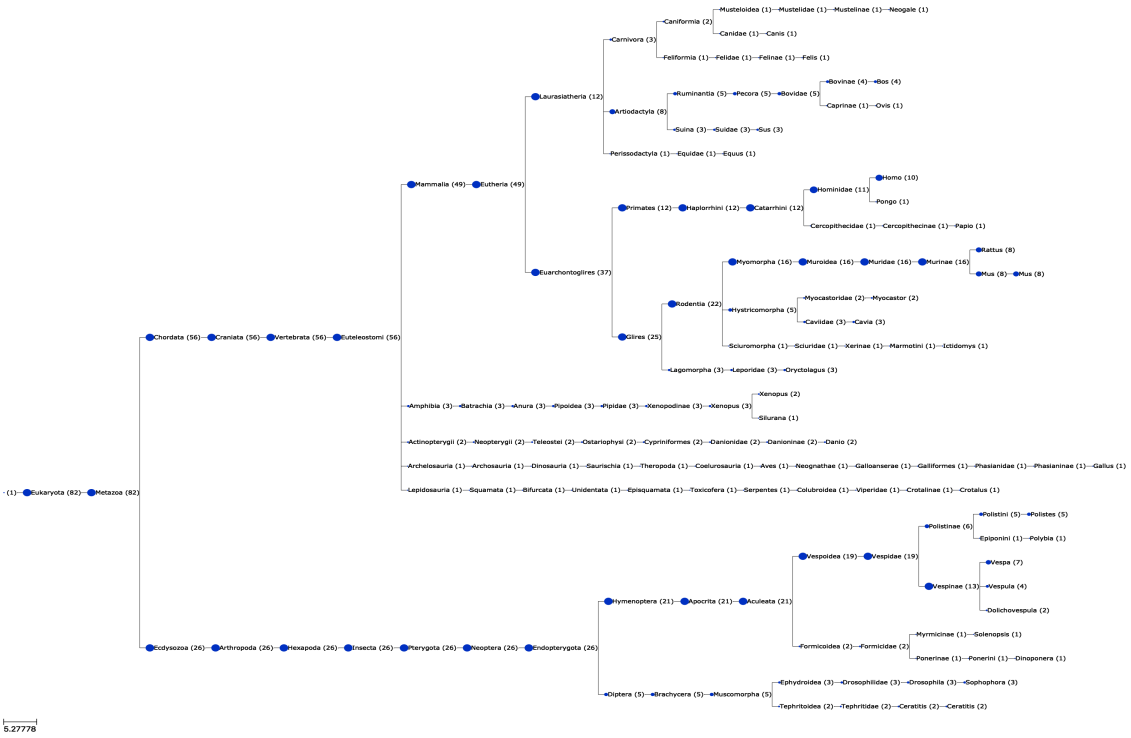

Fig. 2: Phylogenetic tree illustrating taxonomic relationships among family sequences. Node sizes reflect relative abundance, highlighting a dominant presence in mammals.

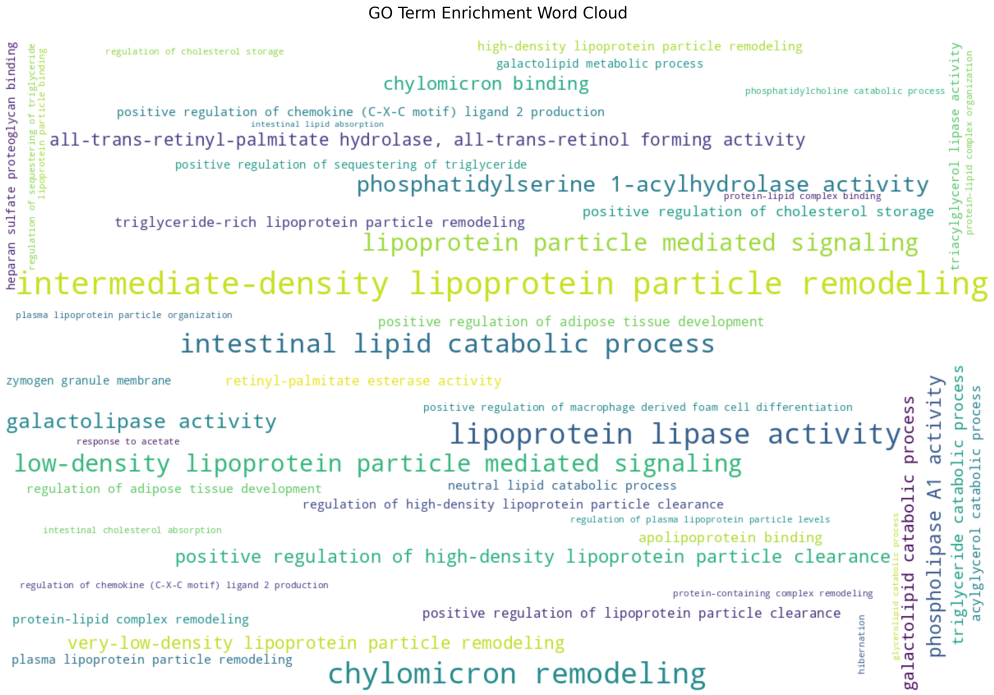

Fig. 3: Word cloud visualization of enriched GO terms. The size of each term represents its relative frequency or significance, with larger text indicating higher enrichment. Terms are related to lipoprotein metabolism, lipase activity, and various cellular processes.

VII. Results

A. Model Evaluation

We evaluated the performance of PSI-BLAST and HMM-SEARCH models at both the protein and residue levels against the SwissProt database. At the protein level, both models successfully identified 82 ground truth proteins, with only one false positive and no false negatives (Table 3). This indicates high precision and recall for both methods. At the residue level, the PSSM model demonstrated a particular high precision and F-score compared to the HMM model, although the HMM model achieved a better recall value (Table 4). The false positive in both models was $P83629$ ($LIPP\_FELCA$). It had the significantly highest e-value and shortest alignment compared to all other hits, which is why we decided to remove it from the family.

TABLE 3: Confusion Matrices for PSI-BLAST and HMM at Protein and Residue Levels

Level	Method	TP	FP	FN	TN
Protein	PSI-BLAST	82	1	0	0
	HMM	82	1	0	0
Residue	PSI-BLAST	22,776	173	2,239	2,428
	HMM	23,287	5,578	1,728	2,280

TABLE 4: Performance Metrics for PSI-BLAST and HMM at Protein and Residue Levels

Metric	Protein Level		Residue Level	
	PSI-BLAST	HMM	PSI-BLAST	HMM
Precision	0.9880	0.9880	0.9925	0.8068
Recall	1.0000	1.0000	0.9105	0.9309
F-score	0.9939	0.9939	0.9497	0.8644
Bal. Acc.	0.5000	0.5000	0.9220	0.6105
MCC	0.0000	0.0000	0.6578	0.2882

The superior performance of the PSSM model can be attributed to its ability to capture position-specific amino acid conservation patterns effectively. However, the HMM model's underperformance may reflect the challenges posed by a relatively small training dataset of 155 sequences and the need for further optimization of its parameters.

B. Taxonomy Analysis

We visualized the taxonomic distribution of the protein family using data from UniProt, plotted as a phylogenetic tree (Fig. 2). The tree revealed a strong focus within *Chordata* ($n = 56$), particularly in *Mammalia* ($n = 49$). The main clusters within mammals include *Rodentia* ($n = 22$) and *Primates* ($n = 12$). These results suggest that the domain is well-conserved across mammalian lineages, which may reflect its fundamental roles in lipid metabolism and energy regulation. The clustering of *Rodentia* and *Primates* further highlights the domain's relevance in key biological functions across diverse species.

C. Gene Ontology Enrichment

Our GO enrichment analysis revealed several significant terms related to lipid metabolism and cellular processes. The most enriched term was *intermediate-density lipoprotein particle remodeling* (Fig. 3), followed by terms such as *chylomicron remodeling* and *intestinal lipid catabolic process*. These terms point to the domain's involvement in lipid processing and dietary fat metabolism. Additional enriched terms included *lipoprotein lipase activity*, which emphasizes the domain's functional role in lipid regulation and enzymatic activity. At the highest ontological level, biological processes showed the strongest enrichment, particularly in cellular and multicellular organismal processes, while molecular functions also displayed significant enrichment, especially in catalytic activity, suggesting broad functional importance across multiple biological scales (Table 5).

GO ID	GO Term	Depth	Propagated Enrichment
GO:0008150	biological_process	0	6727.26
GO:0003674	molecular_function	0	4983.15
GO:0005575	cellular_component	0	692.51
GO:0032501	multicellular organismal process	1	6727.26
GO:0009987	cellular process	1	6727.26
GO:0003824	catalytic activity	1	4983.15
GO:0065007	biological regulation	1	3669.41
GO:0005488	binding	1	2242.42
GO:0110165	cellular anatomical structure	1	692.51
GO:0050896	response to stimulus	1	611.57

TABLE 5: Enriched Branches and their propagated enrichment values

D. Motif Analysis in Disordered Regions

Our analysis of motifs within disordered regions revealed conserved patterns that may serve as functional hotspots. Out of the 82 proteins analyzed, only seven were found to contain annotated disordered regions (Table 6). From these regions, we identified 134 conserved motifs using ELM and ProSite patterns, with an average of 19.14 motifs per sequence. A deeper analysis of these motifs revealed consistent patterns of specific regulatory elements. We frequently observed $CLV\_NRD\_NRD\_1$ and $CLV\_PCSK\_KEX2\_1$ occurring together within the same proteins. Additionally, MOD_CK2_1 appeared prominently across the proteins. The sequences show particularly high frequencies of serine and lysine residues.

Protein ID	Disordered Regions
P27878	(161, 194), (405, 437)
P02843	(158, 196), (407, 439)
P27587	(158, 191), (399, 422)
P02844	(21, 44), (165, 200), (408, 442)
P06607	(401, 420)
Q3SZ79	(23, 44)
P11602	(471, 490)

TABLE 6: Disordered regions for different protein IDs.

VIII. Conclusion

This study comprehensively characterized the *Lipase/vitellogenin* domain family by integrating PSSM and HMM sequence models, taxonomic distribution analysis, GO term enrichment, and motif identification. These approaches provided a multidimensional perspective on the biological significance of the domain. The PSSM model proved effective in capturing conserved positional patterns, outperforming the HMM model in most metrics due to its simplicity. However, HMM models remain promising with further optimization and larger training datasets. Taxonomic analysis highlighted the domain's evolutionary importance, particularly within *Mammalia*, with notable clusters in *Rodentia* and *Primates*. These results suggest a conserved role in lipid metabolism and energy regulation. GO term enrichment emphasized the domain's central role in lipid processing, with significant terms such as *lipoprotein remodeling* and *lipase activity* highlighting its involvement in critical metabolic pathways. Motif analysis further identified conserved patterns in disordered regions, suggesting their role in protein interactions and regulation. Future efforts should focus on refining computational models and validating the identified motifs experimentally to better understand their molecular mechanisms. In addition, investigating interactions with other domains and pathways could provide a more complete picture of the domain functions.

References

[1] Chris P. Ponting and Robert B. Russell. "The Natural History of Protein Domains". In: *Annual Review of Biophysics and Biomolecular Structure* 31 (2002), pp. 45–71.

[2] Stephen F. Altschul, Thomas L. Madden, and Alejandro A. et al. Schäffer. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25.17 (1997), pp. 3389–3402.

[3] Sean R. Eddy. "Profile hidden Markov models". In: *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] The UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". In: *Nucleic Acids Research* 47.D1 (2019), pp. D506–D515. DOI: 10.1093/nar/gky1049.

[5] Jaina Mistry, Sandra Chuguransky, and et al. "Pfam: The protein families database in 2021". In: *Nucleic Acids Research* 49.D1 (2021), pp. D412–D419. DOI: 10.1093/nar/gkaa913.

[6] Christiam Camacho, George Coulouris, and et al. *BLAST+ command line applications user manual*. https://blast.ncbi.nlm.nih.gov/Blast.cgi. 2009.

[7] Fabian Sievers et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular Systems Biology* 7 (2011), p. 539. DOI: 10.1038/msb.2011.75.

[8] Andrew M. Waterhouse, James B. Procter, and et al. "JalView Version 2–a multiple sequence alignment editor and analysis workbench". In: *Bioinformatics* 25.9 (2009), pp. 1189–1191.

[9] Marc Blum, Hsin-Yu Chang, and et al. "The InterPro protein families and domains database: 20 years on". In: *Nucleic Acids Research* 49.D1 (2021), pp. D344–D354. DOI: 10.1093/nar/gkaa977.

[10] The UniProt Consortium. *UniProtKB API Documentation*. https://www.uniprot.org/api-documentation/uniprotkb. 2025.

[11] Jaime Huerta-Cepas, Francisco Serra, and Peer Bork. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Molecular Biology and Evolution* 33.6 (2016), pp. 1635–1638.

[12] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. *A Perl module for phylogenetic trees in the Newick format*. https://evolution.genetics.washington.edu/phylip/newicktree.html. 2008.

[13] Michael Ashburner, Catherine A. Ball, and et al. "Gene ontology: tool for the unification of biology". In: *Nature Genetics* 25 (2000), pp. 25–29. DOI: 10.1038/75556.

[14] The Gene Ontology Consortium. "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1 (2019), pp. D330–D338. DOI: 10.1093/nar/gky1055.

[15] Ronald A. Fisher. *Statistical Methods for Research Workers*. Fisher's exact test as described in this seminal work is widely used for categorical data analysis. 1925.

[16] Damiano Piovesan, Marco Necci, and et al. "MobiDB 3.0: More annotations for intrinsic disorder, conformational diversity and interactions in proteins". In: *Nucleic Acids Research* 46.D1 (2018), pp. D471–D476. DOI: 10.1093/nar/gkx1071.

[17] Holger Dinkel, Kim Van Roey, and et al. “ELM—the eukaryotic linear motif resource in 2016”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D294–D300. DOI: 10.1093/nar/gkv1291.

[18] Christian J.A. Sigrist, Edouard de Castro, and et al. *Prosite: A protein domain and functional site database*. https://prosite.expasy.org/. 2025.

[19] Stevin Wilson. *PrositePatternsToPythonRegex*. https://github.com/stevin-wilson/PrositePatternsToPythonRegex. 2025.