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 [17]. 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 [1] and an HMM [12] model, which are then evaluated in their performance on the protein and residue level against SwissProt [8] proteins annotated with the 'ground truth' Pfam domain [15] 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 VII.

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 [4] against the UniProt Knowledgebase (UniProtKB) [9] with an e-value threshold of 0.0001. Next, we utilized ClustalOmega [18] 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 [20], resulting in 155 sequences, and then performing a detailed conservation analysis. In particular, for the HMM model input, we first removed columns of residues that had 90% or more gaps, which resulted in roughly removing 70% of the initially 3,254 columns. For the PSSM model input, ... (lorenzo). Given the filtered MSAs, we then proceeded to build both the HMM and PSSM models using HMMER-3.4 and NCBI-BLAST-2.16.0+ [5], 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 HMM-SEARCH and PSI-BLAST with default parameters. Both searches were performed against the manually curated SwissProt database (release 29.11.2024) [8], resulting in

83 predicted proteins with e-values < 0.05 and their 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 [15] utilizing the InterPro API [3]. 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 understand the taxonomic distribution of proteins within our family, we utilized a systematic approach to collect and analyze lineage data for the 83 sequences found in our family. Using the protein identifiers obtained from PSI-BLAST and HMM-SEARCH, we queried the UniProt API [10] to fetch the complete taxonomic lineage for each protein. This process captured the hierarchical classification of each protein within its biological domain, from the broadest category (*Eukaryota*) to the most specific organism classification.The taxonomic tree was plotted using the ETE Toolkit [14], leveraging the Newick tree format [6] we derived from the hierarchical data. 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, showcasing the distribution of proteins across various taxonomic groups.

V. Function

First, we obtain GO annotations [2] 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 [7] 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 [13]. 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).

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

TABLE 2: Contingency table

For each 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 detect enrichment (higher proportion in our family than in SwissProt) and a two-tailed test to detect any significant difference in proportions (either higher or lower). The enrichment value was then calculated as:

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

We generated a word cloud visualization using the enriched 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 reported the most enriched branches of the ontology tree based on the enriched terms. For each GO term, we parsed the ontology tree [7] up to the root and added the GO term itself as an enriched child to each found ancestor. After this process, we selected only branches - which we defined as the immediate parent of a GO term, or the GO term itself in case of a root term - that had more than 2 enriched children and a maximum depth of 3 to filter for high-level terms. A selection of 10 of these branches, ranked by their cumulative significance score $S = \sum -\log_{10}(p_{\text{two-tailed}})$ calculated across all child terms, can be seen in Table 5.

VI. Motifs

To identify significantly conserved short motifs within our protein family, we first needed to determine the locations of intrinsically disordered regions. These regions were extracted from MobiDB-lite [16], a database that employs a consensus-based approach to predict protein disorder. Subsequently, we extracted motif patterns from ELM classes [11] and ProSite patterns [19] and searched for these patterns within the previously identified disordered regions. To convert ProSite patterns into Python compatible regular expressions, we used a script adapted from Stevin Wilson's $PrositePatternsToPythonRegex$ project [21].

VII. Results

A. Model Evaluation

The models were evaluated on both the protein and residue levels using PSI-BLAST and HMM-SEARCH against the SwissProt database. Key findings include:

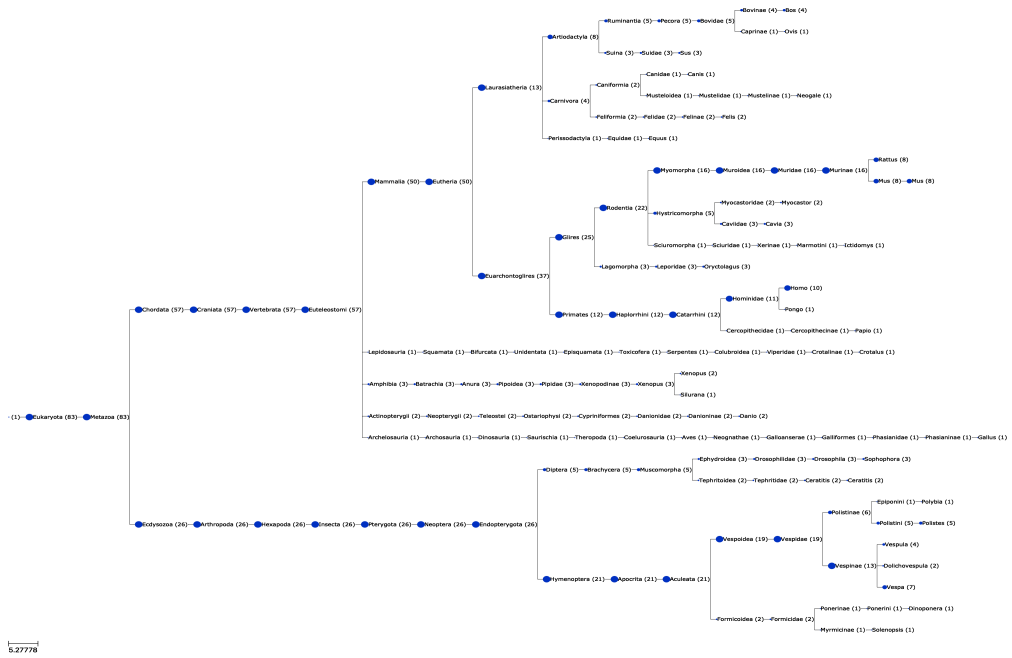

Fig. 2: Phylogenetic tree illustrating taxonomic relationships across family sequences, with node sizes reflecting relative abundance. The tree highlights a dominant presence in mammals, particularly *Primates*, and significant diversity within arthropods. This visualization underscores the evolutionary adaptation and taxonomic distribution of the protein domain.

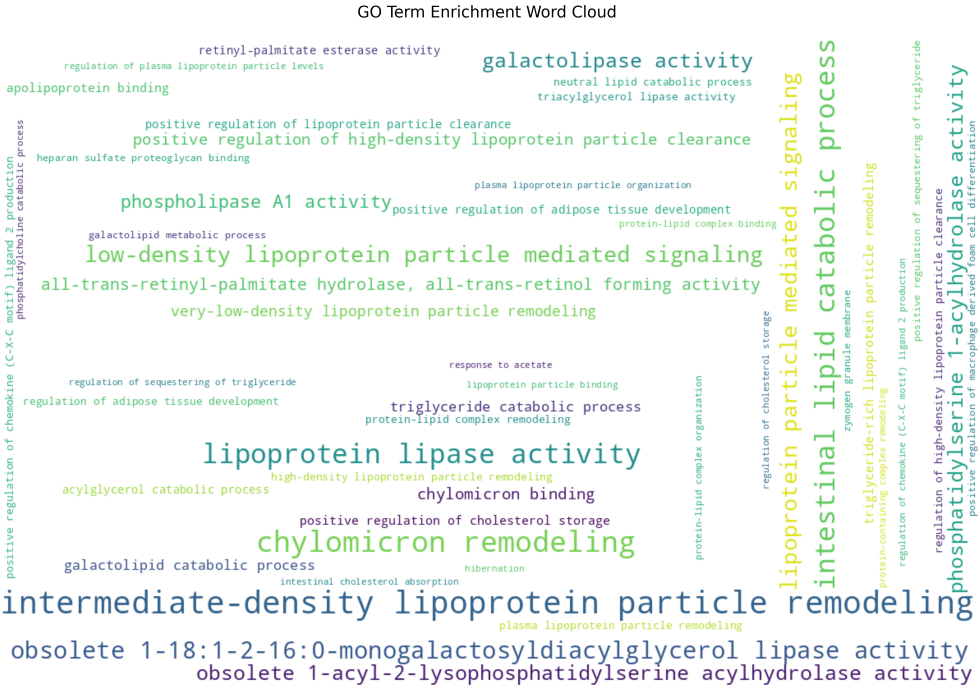

Fig. 3: Word cloud visualization of enriched GO terms. The size of each term represents its relative frequency or significance, with larger text indicating higher enrichment. Terms are related to lipoprotein metabolism, lipase activity, and various cellular processes.

- **Protein-Level Metrics:** Both models identified 82 ground truth proteins with only one false positive and no false negatives, indicating high precision and recall (Table 3).
- **Residue-Level Metrics:** The PSSM model outperformed the HMM model in every metric. However, it is notable that the HMM model demonstrated a competitive recall score. (Table 4).

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	23,341	2,157	1,674	2,357
	HMM	23,322	5,580	1,693	2,277

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	0.9154	0.8069
Recall	1.0000	1.0000	0.9331	0.9323
F-score	0.9939	0.9939	0.9242	0.8651
Bal. Acc.	0.5000	0.5000	0.7276	0.6111
MCC	0.0000	0.0000	0.4772	0.2907

The superior performance of the PSSM model, despite its simpler architecture compared to HMM, may be explained by several factors. First, the domain family likely exhibits strong position-specific amino acid conservation patterns that PSSM models are particularly adept at capturing. Second, our relatively small training dataset (i.e. 155 sequences) may be insufficient to properly estimate the numerous parameters required by the more complex HMM model. Additionally, the HMM performance might be improved through optimization of its search parameters, as we utilized default settings in our analysis.

B. Taxonomy Analysis

Using lineage data from UniProt, we visualized the taxonomic distribution of the family proteins. The phylogenetic tree (Fig. 2) reveals a distribution primarily focused within *Chordata* ($n = 57$), with a strong representation in *Mammalia* ($n = 50$). Within mammals, the domain shows significant presence across multiple orders, particularly in:

- *Rodentia* ($n = 22$), which includes mice and rats
- *Primates* ($n = 12$)

The strong conservation across mammalian orders, from *Rodentia* to *Primates*, indicates the domain likely plays a fundamental role that has been maintained throughout mammalian evolution.

C. Gene Ontology Enrichment

The functional enrichment analysis identified several highly significant GO terms, with intermediate-density lipoprotein particle remodeling emerging as the most enriched function (Fig. 3). Closely following was chylomicron remodeling, which together with intestinal lipid catabolic process suggests a crucial role in dietary fat processing. The presence of lipoprotein lipase activity among the top terms reinforces this domain's importance in lipid metabolism. Multiple related signaling functions were identified with identical enrichment weights , including both general lipoprotein particle mediated signaling and the more specific low-density lipoprotein particle mediated signaling. The domain also shows strong association with various lipase activities, including galactolipase activity. This enrichment pattern reveals a domain family specialized in lipoprotein modification and lipid processing, with particular emphasis on the metabolism of dietary fats. The Analysis of the enriched GO branches reveals a clear hierarchical organization of function (Table 5). The biological process root category shows the highest enrichment , followed by molecular function, indicating broad involvement across cellular processes. The strong representation of metabolic processes and their subcategories validates our earlier findings about the domain family's specialization in dietary fat metabolism. Particularly noteworthy is the significant enrichment of catalytic activities, specifically hydrolase activity and its more specialized child term 'hydrolase activity, acting on ester bonds'. This hierarchical pattern of enrichment further supports the domain's crucial role in lipid processing through specific enzymatic activities.

GO Term	Branch Name	Depth	Enriched	S
biological_process	biological_process	0	212	2654.64
cellular process	biological_process	1	83	1685.44
molecular_function	molecular_function	0	42	1407.35
metabolic process	cellular process	2	58	1161.52
catalytic activity	molecular_function	1	20	1066.89
hydrolase activity	catalytic activity	2	17	1006.90
primary metabolic process	metabolic process	3	40	993.51
hydrolase activity, acting on ester bonds	hydrolase activity	3	15	916.11
biological regulation	biological_process	1	68	529.29
regulation of biological process	biological regulation	2	64	515.90

TABLE 5: Top 10 enriched GO terms and their branches

D. Motif Analysis in Disordered Regions

During our analysis of Motifs, we found (Table 6):

- Only 7 out of 83 family proteins had annotated disordered regions (Table 6).
- Motif analysis using ELM and ProSite patterns identified 134 conserved motifs, with an average of 19.14 motifs per sequence. This indicates potential functional hotspots within the disordered regions.

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

In this study, we delved into the Lipase/vitellogenin domain family. The combination of PSSM and HMM modeling approaches, along with detailed taxonomic distribution and GO term enrichment analysis, has helped establish a robust framework for understanding this domain's role across different organisms. Looking forward, several promising directions emerge. Refining our models by optimizing gap penalties during the MSA creation or adding additional sequence data can further improve their performance, the latter especially in the case of the HMM model, which lacks performance as of now. A deeper investigation of the taxonomic distribution could reveal interesting patterns of domain loss or gain across different lineages, potentially showcasing the domain's evolutionary history. Lastly, exploring how this domain family interacts with other domains could provide a more complete picture of its biological role and evolution.

References

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[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] 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.

[18] 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.

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

[20] 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.

[21] Stevin Wilson. *Prosite Patterns To Python Regex*. https : / / github . com / stevin - wilson / PrositePatternsToPythonRegex. 2025.