Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Alessia Corica[1]

[1] Bioinformatics Master's Degree Course, University of Bologna, Italy

For correspondence: alessia.corica@studio.unibo.it

Laboratory of Bioinformatics I - Module 2, Academic year 2024/2025

Final submission: May 2025

Abstract

Motivation: The Kunitz-type protease inhibitor domain is a structurally conserved motif found in a wide variety of proteins involved in regulatory and inhibitory functions. This project aimed to build a profile Hidden Markov Model (HMM) for the identification of Kunitz domains. The model was constructed using a structure-based multiple sequence alignment and validated on curated positive and negative datasets, with the goal of enhancing the accuracy of automated domain annotation.

Results: Cross-validation confirmed the high classification performance of the structure-based HMM. At the optimal E-value threshold of $1e^{-9}$, the model achieved an MCC of 1.000 in Fold 1 and 0.9918 in Fold 2. Precision, recall, and accuracy exceeded 0.98 in both cases, with a perfect classification in Fold 1 (0 false positives, 0 false negatives) and minimal misclassification in Fold 2 (2 false positives, 1 false negative). These results support the model's robustness and specificity in identifying Kunitz domains.
Structural analysis of the few misclassified sequences suggested they retained canonical Kunitz features, pointing to potential annotation issues in the UniProtKB/SwissProt reference database.

Supplementary information: All figures, scripts, raw and output files mentioned in this study are available in the project's GitHub repository https://github.com/alessia-corica/LAB1_Kunitz_project

1. Introduction

1.1.Biological role of Kunitz proteins domain

The Kunitz domain is a widely conserved functional unit found in a variety of proteins across many animal species. It plays a central role in inhibiting serine proteases, thereby contributing to the regulation of critical physiological processes such as blood coagulation, fibrinolysis, inflammation, and host immune defense (Ranasinghe & McManus, 2013).
Bovine pancreatic trypsin inhibitor (BPTI) was the first Kunitz-type protease inhibitor described (Kunitz and Northrop, 1936). Due to its well-characterized inhibitory activity and broad scientific relevance, BPTI has long served as a model system in structural, biochemical, and computational studies of serine protease inhibition. Given its functional relevance and evolutionary conservation, the Kunitz domain is a valuable target for computational domain detection and sequence-based annotation tools.

1.2. Structural Features and Stability

The Kunitz domain exhibits a α+β fold, stabilized by three highly conserved disulfide bridges arranged in a C1–C6, C2–C4, and C3–C5 configuration. These three covalent linkages play a critical role in maintaining the structural integrity

and stability of the domain, even under proteolytic stress (Ranasinghe & McManus, 2013).

Another functional feature is the presence of a positively charged residue at position 15, typically lysine or arginine, which serves as the reactive site for serine protease inhibition by occupying the substrate-binding pocket (Rawlings et al., 2004).

As shown in Figure 1, the bovine pancreatic trypsin inhibitor (BPTI) displays these hallmarks of the Kunitz fold, including the disulfide pattern, secondary structure arrangement, and the insertion of Lys15 into the active site of the target enzyme.

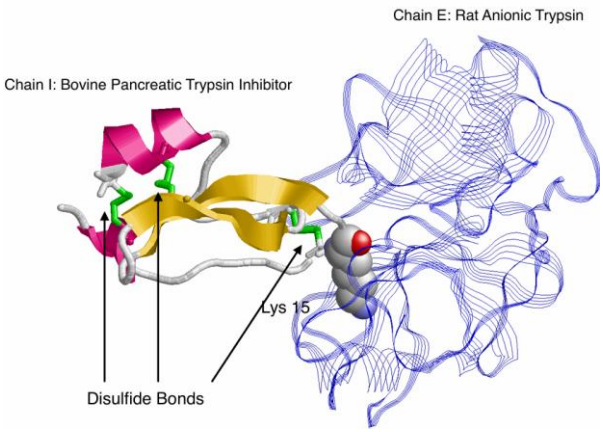

Figure 1. Structure of the Kunitz domain in BPTI (Chain I) in complex with rat anionic trypsin (Chain E)

1.3. Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) are probabilistic models used to represent biological sequence families by capturing both conserved and variable regions. In a profile HMM, each position in a multiple sequence alignment is modeled by a set of hidden states (match, insert, and delete) that emit or handle residues with specific probabilities.

Given a trained model, a new sequence can be scored based on how likely it is to be generated by the model, enabling sensitive detection of domain instances even in divergent sequences. This makes HMMs especially suited for identifying conserved domains such as the Kunitz-type protease inhibitor, where structural constraints result in characteristic sequence signatures.

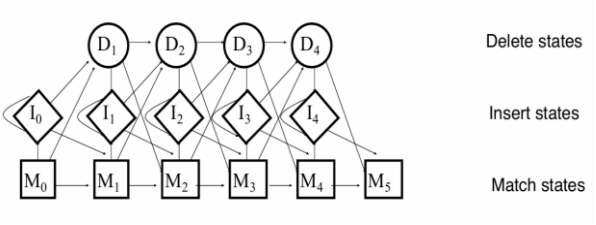

Figure 2. Schematic representation of a profile Hidden Markov Model (HMM), showing match (M), insert (I), and delete (D) states and their possible transitions.

1.4.Objectives and workflow of this study

The main objective of this study is to build a profile HMM specifically tailored to detect the Kunitz-type protease inhibitor domain, starting from structurally curated alignments. The model was trained using a multiple sequence alignment derived from structural superposition of experimentally resolved proteins. This approach helps ensure that functionally important residues, such as cysteines forming disulfide bonds, are consistently aligned across sequences, enhancing the biological relevance of the HMM.

The workflow begins with the selection and clustering of high-resolution protein structures from the PDB annotated with the PFAM domain PF00014. A structure-based alignment is generated and converted into a multiple sequence alignment, which is then used to train the profile HMM using HMMER (Eddy, 2011).

The model is validated on a curated dataset of UniProt/SwissProt sequences, categorized into positive and negative sets. Performance is assessed using metrics such as Matthews Correlation Coefficient (MCC), F1 score, and ROC analysis. Particular attention is given to borderline predictions and annotation mismatches, in order to evaluate whether structurally informed models perform better than those based on sequence alone.

2. Materials and Methods

2.1. Data collection

The reference database used for constructing the structural model was the Protein Data Bank (PDB) (Berman et al.,2000). To ensure domain-specificity and structural quality, an advanced query was performed using the filtering parameters listed in Table 1.

Filter	Criteria
PFAM Domain Identifier	PF00014
Data Collection Resolution	≤ 3.0 Å
Polymer Entity Sequence Length	> 50 and ≤ 80 residues

Table 1. Filters applied for structural data retrieval from PDB.

The final set of representative structures contained a list of 158 entries and, that was exported as a custom csv report and downloaded as rcsb_pdb_custom_report_20250505025420.csv, which is available in the project's GitHub repository (*LAB1_Kunitz_project/raw_data/*). This file contains, for each entry, the PDB ID, amino acid sequence, chain identifier, PFAM domain annotation, and Gene Ontology (GO) terms.

The structural sequences annotated with the Kunitz domain were converted into FASTA format. The resulting file, pdb_kunitz_customreported.fasta/*processed_data*/ , was clustered at 90% sequence identity using CD-HIT (Fu et al., 2012) to remove redundancy. One outlier (2ODY_E) was excluded manually due to its atypical sequence length and the final set of 24 representative sequences was saved as a FASTA file (pdb_kunitz_rp.fasta/*processed_data/*) for alignment and HMM training.

A complete set of 395 reviewed UniProt proteins annotated with the Kunitz domain was downloaded from the UniProt database (The UniProt Consortium, 2023) and stored in the file kunitz_sequences.fasta/*raw_data*/. The dataset was subsequently divided into two subsets: 18 human sequences (human_kunitz_sequences.fasta/*raw_data*/) and 377 non-human sequences (nothuman_kunitz_sequences.fasta/*_raw_data/*).

2.2. Model building

To prepare the input for structural alignment, the file pdb_kunitz_rp.fasta was reformatted to comply with the PDBeFold (Krissinel & Henrick, 2004) input requirements. Specifically, each entry was edited to include only the PDB identifier and chain ID (e.g., 5NX1:C), without any sequence data. This minimal list was saved in plain text format (tmp_pdb_efold_ids.txt/*processed_data/*) and uploaded to the PDBeFold server for batch structural alignment. One aligned entry was manually removed due to its low number of aligned residues and high RMSD, indicating poor structural conservation with respect to the core Kunitz fold. As a result, 23 structures were retained for model construction.

RMSD values from PDBeFold alignment are shown in Figure 3 to illustrate the structural compatibility of each entry with the reference fold. A threshold of 1.0 Å was used as a qualitative reference to highlight structurally well-aligned entries, as RMSD values below this level are generally considered indicative of strong structural conservation in domain-level comparisons.

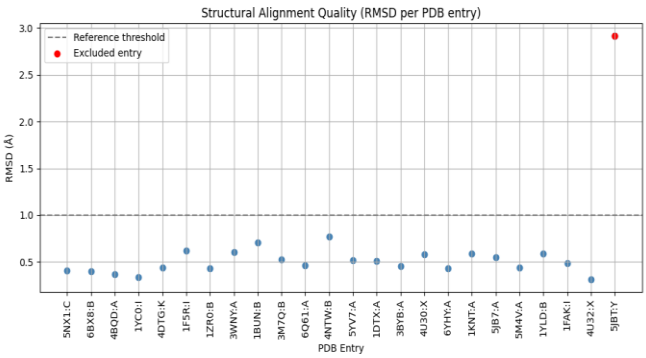

Figure 3. RMSD values from PDBeFold alignment across all PDB entries. The red point marks the excluded structure (5JBT:Y) due to poor alignment.

The resulting alignment was used as input for the hmmbuild program from the HMMER v3.3.2 suite (Eddy, 2011) to construct the profile Hidden Markov Model. The resulting model was saved as structural_model.hmm and stored in the

hmm_model/ directory to be used for all subsequent hmmsearch analyses.
The resulting profile reflects the conserved features of the Kunitz domain across structurally aligned sequences.
To visualize the level of residue conservation across the alignment, sequence logos were generated using both WebLogo and Skylign (Wheeler et al., 2014; Crooks et al., 2004).

2.3. Generation of positive and negative sets

To assess the model's ability to distinguish true Kunitz domain sequences from unrelated ones, two distinct benchmark datasets were constructed from UniProt/SwissProt entries annotated with the Kunitz domain (PF00014).
First, redundancy in the structural dataset was eliminated through a BLAST search that was performed by aligning the 23 representative structural sequences against the full Kunitz set using blastp. The BLAST alignment was saved as pdb_kunitz_23.blast/*results*/*blast*/ and filtered to retain only hits with sequence identity ≥95% and alignment coverage ≥50%, which were considered redundant.
The UniProt IDs corresponding to these redundant sequences were identified and subtracted from the full Kunitz list. The remaining 366 non-redundant sequences were selected to form the final positive set and stored in ok_kunitz.fasta/*results*/*blast*/.
The negative set was generated from the SwissProt database by extracting reviewed protein sequences not annotated with the Kunitz domain (PF00014), for a total of 572,835 sequences that were assumed to lack Kunitz-like structural or functional features.
To ensure a balanced evaluation, both positive and negative datasets were randomly shuffled and evenly split into two subsets using shell commands, maintaining reproducibility and balanced class distribution across folds. Sequence extraction from the reference FASTA files was performed using a custom Python script (get_seq.py/*scripts*/), and the resulting files were organized into two cross-validation folds, each comprising a positive and a negative set.

2.4. Model testing

The obtained profile HMM was tested on the benchmark subsets using hmmsearch (Eddy, 2011). Each of the four FASTA subsets was queried independently to evaluate classification performance across both folds.
The --max and -Z 1000 options were used to normalize E-values and improve sensitivity. In particular, --max disables heuristic filters and perform full alignment on all hits, allowing weaker matches to be retained. This helps detect more true positives, increasing the sensitivity of the search. Meanwhile, -Z 1000 adjusts the E-value calculation to make results more comparable across datasets of different sizes.
The output tables were then parsed to extract the UniProt ID, the E-value and a binary label: 1 for sequences from the positive sets and 0 for negatives. The resulting classification tables were used for downstream threshold-based evaluation.

2.5. Performance evaluation

Classification performance was evaluated by applying E-value thresholds to the hmmsearch output and computing standard metrics: Matthews Correlation Coefficient (MCC), accuracy, precision, and recall, based on the classification of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

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

MCC is a metric that measures the overall quality of a binary classification.

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

Accuracy reflects the proportion of correctly classified instances.

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Precision measures how many of the predicted positives are actual positives.

$$Recall = \frac{TP}{TP + FN}$$

Recall (also known as sensitivity or true positive rate) indicates the proportion of true positives identified by the model.

A fixed E-value cutoff of 1e-5 was initially used as a standard threshold, as commonly adopted in HMM-based classification to balance sensitivity and specificity. This value served as a baseline for identifying borderline cases. Sequences from the positive sets that scored above this threshold were considered false negatives and stored in fn_pos1.txt and fn_pos2.txt/*results/errors/*).
For the negative sets, sequences that scored below the threshold were classified as false positives. However, only set 2 produced such cases, which were stored in fp_neg2.txt/*results/errors/*).
A broader evaluation was then performed by testing ten E-value thresholds, exponentially spaced from 1e-10 to 1e-1, using the script performance.py/*scripts/*) to compute metrics for each threshold.
Based on this analysis, the threshold yielding the highest average MCC (1e-9) was selected for final evaluation. Classification metrics (MCC, precision, recall, accuracy) were recalculated at this optimal cutoff using both benchmark folds.

2.6. Qualitative structural validation

To investigate the structural characteristics of misclassified sequences, both false negatives and false positives identified at the fixed E-value threshold of 1e-5 were subjected to additional analysis.
Two sequences for each category were selected for structural comparison. Their AlphaFold-predicted models were aligned with the reference structure of BPTI (PDB ID: 3TGI) using UCSF ChimeraX (Pettersen et al., 2021).
This qualitative analysis aimed to determine whether misclassified sequences retain structural features characteristic of the Kunitz domain. The goal was to assess if certain borderline cases were due to model limitations or inconsistencies in domain annotation. By comparing predicted structures with a known reference, this step aimed to complement the quantitative evaluation with structural insight.

3. Results and Discussion

3.*1. Classification performance at a fixed threshold*

The model was first evaluated using a fixed E-value threshold of 1e-5, a standard cutoff often used in HMM-based classification. At this threshold, the classifier performed very well on both validation folds, correctly identifying most Kunitz sequences and minimizing misclassifications. These results indicate that the model achieves a good balance between sensitivity and specificity under standard conditions.
To assess whether performance could improve further, we repeated the evaluation using the optimal threshold of 1e-9, selected based on the highest average MCC across thresholds. As shown in Tables 2 and 3, the results at 1e-9 were nearly identical to those at 1e-5. This confirms that the model is robust to changes in the threshold and maintains high performance even with more stringent criteria.
Overall, the comparison supports the use of 1e-5 as a reliable default, while 1e-9 may be preferred when prioritizing strict classification and minimizing borderline cases.

Fold	TP	TN	FP	FN	MCC	Precision	Recall	Accuracy
1	180	286417	0	3	0.9918	1.000	0.9836	1.000
2	182	286415	2	1	0.9918	0.9891	0.9945	1.000

Table 2. Classification performance at E-value threshold 1e-5

Fold	TP	TN	FP	FN	MCC	Precision	Recall	Accuracy
1	183	286417	0	0	1.0000	1.000	1.000	1.000
2	182	286415	2	1	0.9918	0.9891	0.9945	1.000

Table 3. Classification performance at E-value threshold 1e-9

In addition to the overall performance metrics, confusion matrices were generated to visualize the distribution of predictions for each fold. This representation allows for a clearer understanding of

how the model separates positive and negative cases and helps to immediately identify the presence of false positives or false negatives. The matrices complement the numerical results by providing an intuitive summary of classification outcomes at both thresholds.

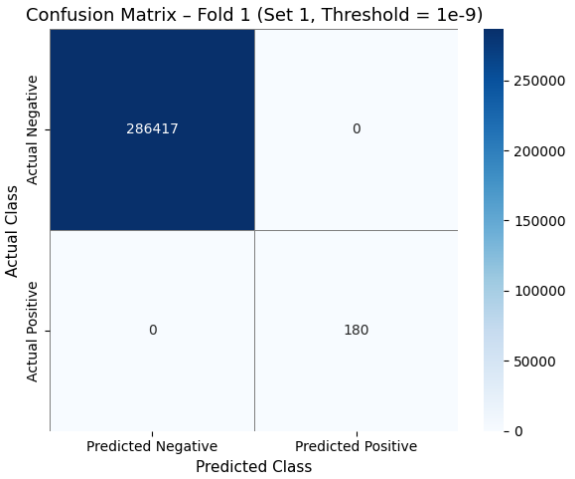

Figure 4. Confusion matrix for Fold 1 at threshold 1e-9. The model correctly classified all sequences with no false positives or false negatives.

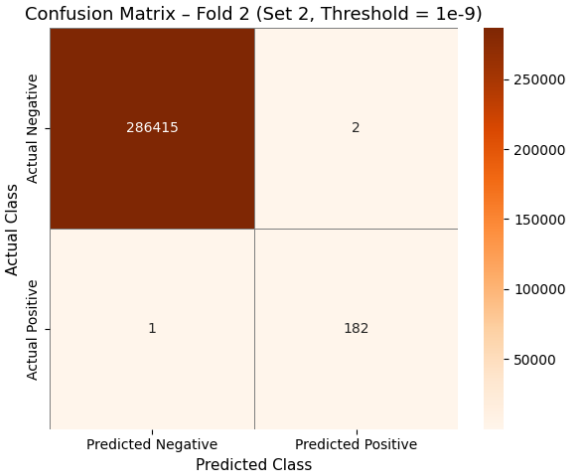

Figure 5. Confusion matrix for Fold 2 at threshold 1e-9. The model produced only 3 misclassifications.

3.1. *Classification performance across 10 thresholds*

To evaluate the robustness of the model and identify the optimal classification threshold, we evaluated 10 E-value cutoffs ranging from 1e-10 to 1e-1. For each value, we calculated performance metrics on both validation folds using the same classification results used for the classification at fixed thresholds.

Among the evaluated metrics, MCC was chosen for visualization as it provides a balanced measure even in the presence of class imbalance, making it particularly suited for this dataset.

As shown in Figure 6, the MCC remained consistently high from 1e-10 to 1e-5, indicating that the model is stable and performs well across a wide range of thresholds. The best average performance was observed at 1e-9, where both folds achieved near-perfect scores.

Performance started to drop at more relaxed thresholds (≥ 1e-3), particularly for Fold 2. This suggests that stricter cutoffs help reduce false positives in borderline cases.

Overall, these results confirm that 1e-5 is a solid standard, while 1e-9 may be preferred for more conservative classification.

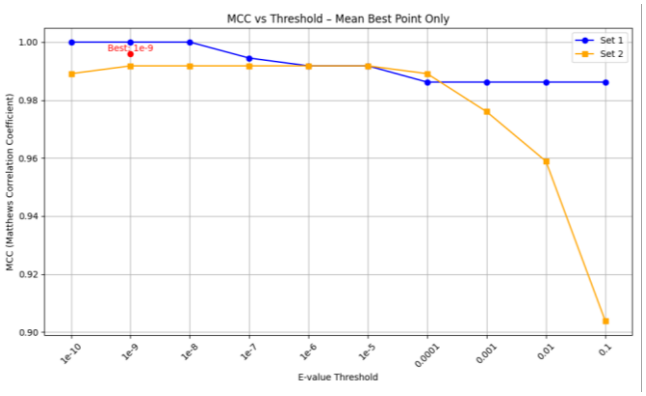

Figure 6. MCC scores across ten E-value thresholds. The highest average MCC was obtained at 1e-9, with performance remaining high up to 1e-5.

3.3. *Analysis of misclassified sequences*

To better understand the classification errors, we performed a qualitative structural comparison using UCSF ChimeraX (Pettersen et al., 2021). Two false negatives and two false positives, identified at the 1e-5 threshold, were selected and aligned to the reference structure of BPTI (PDB ID: 3TGI) based on their AlphaFold-predicted models. As shown in Figure 7, both false negatives (AF-D3GGZ8-F1 and AF-Q8WPG5-F1) showed a structural overlap with the reference domain, despite being misclassified by the model as non-Kunitz. This suggests that the model may have

failed to recognize the domain, maybe because of extended or variable regions affecting the E-value, leading to incorrect predictions even though the core Kunitz fold is preserved.

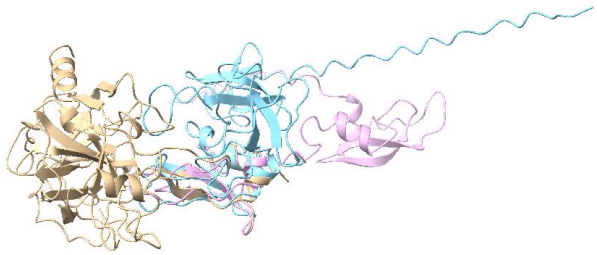

Figure 7. Structural alignment of false negatives (AF-D3GGZ8-F1 in light blue and AF-Q8WPG5-F1 in pink) to the reference Kunitz domain (3TGI in beige).

Interestingly, in Figure 8 it's possible to see that the structural superimposition of the two false positives (AF-P0DQR0-F1 and AF-P0DQQ9-F1) also revealed high structural similarity with BPTI, particularly in the conserved β-sheet and α-helix arrangement typical of the Kunitz domain. Further inspection confirmed that both proteins are annotated in UniProt as "Kunitz-like" or "Kunitz-type protease inhibitors," reinforcing the idea that these are not true misclassifications. Instead, they likely reflect annotation discrepancies in the reference dataset.

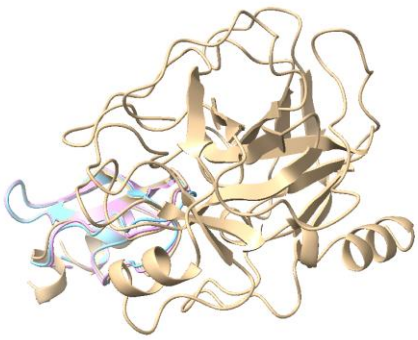

Figure 8. Structural alignment of false positives (AF-P0DQR0-F1 in light blue and AF-P0DQQ9-F1 in pink) to the reference Kunitz domain (3TGI, beige).

Overall, this structural validation supports the model's ability to capture functionally relevant features and highlights the limitations of relying solely on database annotations.

3.4. ROC curve analysis

To further evaluate the model's ability to discriminate between Kunitz and non-Kunitz sequences, the Receiver Operating Characteristic (ROC) curve was computed based on the based on the classification results across a wide range of E-value thresholds. The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (FPR), providing an overall measure of classification performance independent of a specific cutoff.

$$\boldsymbol{TPR = \frac{TP}{TP + FN}} \qquad \boldsymbol{FPR = \frac{FP}{FP + TN}}$$

As shown in Figure 9, the curve closely approaches the top-left corner of the plot, indicating high sensitivity and low false positive rate across most thresholds. This results in a nearly perfect Area Under the Curve (AUC), confirming the model's strong ability to separate positive and negative cases.

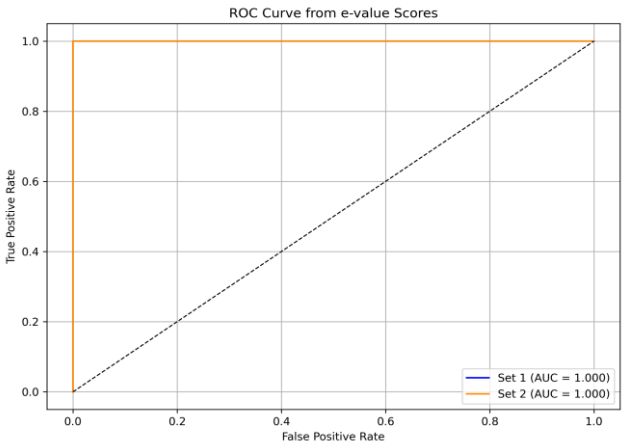

Figure 9. ROC curve of the HMM classifier.

This analysis supports the results obtained at fixed cutoffs, supporting the robustness and consistency of the classifier across different levels of stringency.

3.5. Sequence logo analysis

To explore the conservation patterns across the aligned sequences, sequence logos were generated

from the structural alignment using Skylign and WebLogo (Crooks et al., 2004; Wheeler et al., 2014). As shown in Figures 10 and 11, both representations highlight highly conserved residues within the Kunitz domain, particularly at cysteine positions involved in disulfide bond formation. These residues are functionally relevant for the structural stability and inhibitory function of the domain. The consistency between the two logos confirms that the structure-based alignment effectively captures the core conserved features, supporting the biological relevance of the HMM profile.

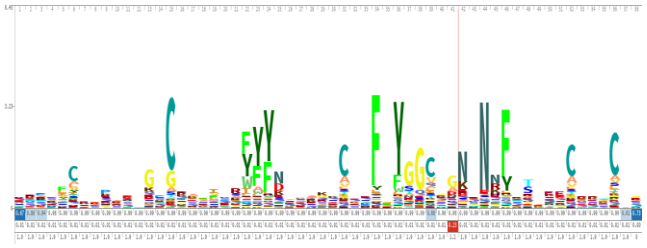

Figure 10. Sequence logo generated with Skylign. Conserved cysteine residues and functional motifs are clearly highlighted.

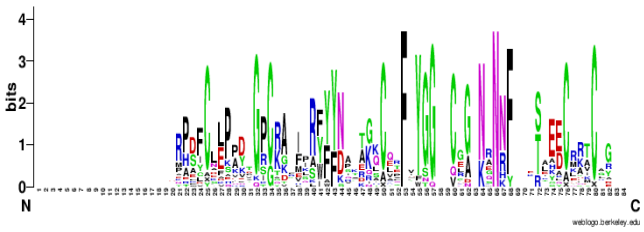

Figure 11. Sequence logo generated with WebLogo, showing conservation across the full domain profile.

Conclusions

This study presented a structure-informed approach for detecting Kunitz domains using profile Hidden Markov Models (HMMs). By leveraging multiple sequence alignment derived from experimentally resolved protein structures, we generated a robust profile HMM that captured key conserved features of the domain. Evaluation on curated benchmark datasets demonstrated high accuracy, precision, and recall across both validation folds, confirming the model's effectiveness in identifying true Kunitz-type protease inhibitors.

Threshold-based classification and ROC curve analysis further validated the model's performance, with the optimal E-value threshold identified at 1e-9. At this cutoff, the model achieved near-perfect scores, highlighting its ability to maintain high sensitivity while minimizing false positives. Structural superimposition of selected false negatives and false positives revealed strong overlap with the reference fold, supporting the idea that some apparent misclassifications were likely due to annotation inconsistencies in the reference dataset.

The sequence logos generated from both the structure-based and profile-based alignments confirmed the conservation of functionally relevant residues, such as cysteines involved in disulfide bonding. This conservation strengthens confidence in the alignment quality and the biological relevance of the resulting HMM.

Overall, this pipeline demonstrates the value of integrating structural information into domain annotation workflows. It enhances the specificity and sensitivity of domain detection, particularly in borderline or misannotated cases.

Future directions could include extending this approach to other small, disulfide-rich protein families, such as defensins or toxin-like peptides, which often display high structural conservation despite low sequence identity. Additionally, integrating functional data or structural confidence scores (e.g., AlphaFold pLDDT) may further improve discrimination in ambiguous cases. Applying the model to large-scale proteome annotation or metagenomic datasets could also reveal novel Kunitz-like domains currently lacking functional annotation.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*

Res, **25**(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**(1):235–242. https://doi.org/10.1093/nar/28.1.235

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**(6):1188–1190. https://doi.org/10.1101/gr.849004

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol*, **7**(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, **60**(12):2256–2268. https://doi.org/10.1107/S0907444904026460

Kunitz M, Northrop JH (1936) Crystalline trypsinogen and its conversion to crystalline trypsin. *J Gen Physiol*, **19**(6):991–1007. https://doi.org/10.1085/jgp.19.6.991

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*, **30**(1):70–82. https://doi.org/10.1002/pro.3943

Ranasinghe S, McManus DP (2013) Structure and function of invertebrate Kunitz serine protease inhibitors. *Dev Comp Immunol*, **39**(3):219–227. https://doi.org/10.1016/j.dci.2012.10.005

Rawlings ND, Tolle DP, Barrett AJ (2004) Evolutionary families of peptidase inhibitors. *Biochem J*, **378**(3):705–716. https://doi.org/10.1042/BJ20031825

The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*, **51**(D1):D523–D531. https://doi.org/10.1093/nar/gkac1052

Wheeler TJ, Clements J, Finn RD (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, **15**:7. https://doi.org/10.1186/1471-2105-15-7