A Bioinformatics Workflow for Modeling and Detecting the Kunitz Domain Using Profile HMMs

Bianca Mastroddi

Department of Pharmacy and Biotechnology, University of Bologna, Italy.
Bioinformatics Master's Degree Course

May 2025

Supplementary materials: https://github.com/bianca29092001/kunitz-domain

Abstract. Kunitz domains are small structural motifs that act as serine protease inhibitors and are involved in key biological processes such as coagulation, inflammation, and neural regulation. This project aims to develop a Profile Hidden Markov Model (HMM) specifically tailored for the Kunitz domain by leveraging structural information extracted from high-resolution PDB entries. Starting from a curated set of 23 representative protein structures, we built a multiple structural alignment and converted it into a sequence alignment used to train the HMM. Benchmark datasets were generated from SwissProt and refined to remove sequences overly similar to the training set. The model was tested via 2-fold cross-validation and evaluated using metrics such as accuracy and Matthews Correlation Coefficient (MCC), reaching values above 0.99 and 0.89 respectively. The method demonstrated robust and generalizable predictions, even when applied to non-human proteins. This work highlights the power of combining structural and sequence data for reliable domain annotation, and sets a framework for future extension to other protein families.

Introduction

Proteins containing the Kunitz domain play a crucial role in modulating the activity of serine proteases, often acting as inhibitors in processes where unregulated proteolysis would be harmful. These domains are typically short—around 60 amino acids—but structurally stable, owing to the presence of conserved cysteines forming disulfide bridges [1]. Their inhibitory function and structural rigidity make them widespread in diverse species and relevant across physiological and pathological contexts.

In humans, Kunitz-containing proteins are implicated in blood coagulation (e.g., TFPI), neurodegeneration (e.g., APP), and immune responses, while in other organisms they participate in mechanisms such as venom

toxicity or parasite defense [2]. Due to their functional significance, accurate detection of Kunitz domains is essential. Misidentification or missed detection of these domains can lead to incomplete or incorrect protein function annotation, particularly in large-scale genomic or proteomic projects. Clinically, this may result in overlooking potential therapeutic targets or biomarkers.

From a computational perspective, detecting protein domains based solely on sequence similarity can be challenging—especially when evolutionary divergence leads to low sequence identity but conserved 3D structures. As a result, models that incorporate structural data are becoming increasingly important for reliable domain prediction. [3]

This project focuses on building a Hidden Markov Model (HMM) derived from structure-based multiple alignments, rather than conventional sequence-only approaches [4].

By using a training set of non-redundant protein structures annotated with the Kunitz domain, we aim to build a model that captures conserved biochemical and spatial features.

The trained model is then validated on benchmark datasets to assess its ability to discriminate between true and false domain instances, using standardized performance metrics. This approach offers a refined and evolutionarily aware tool for domain annotation, with the potential to be extended to other protein families in future work.

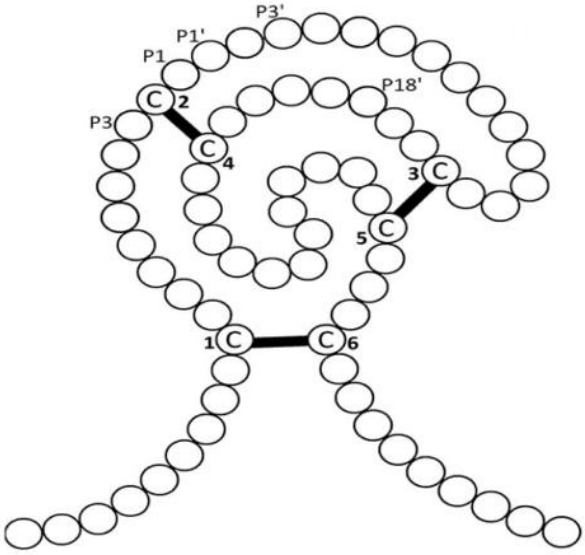

Figure 1. Schematic of the Kunitz domain fold, stabilized by three disulfide bonds formed by six conserved cysteines. These structural features are essential for maintaining the domain's conformation and enabling its inhibitory interaction with serine proteases

Materials and methods

2.1 Dataset Generation

A curated set of PDB structures containing the Kunitz domain was downloaded based on resolution (≤ 3.5 Å), domain presence (PF00014), and chain length (45-80 amino acids). Only reviewed entries were considered. Sequences were filtered to retain non-redundant representatives using CD-HIT at 90% sequence identity, followed by manual filtering to remove overly long sequences. Representative sequences were selected based on cluster centroids, resulting in 23 structures for alignment.

2.2 Multiple Structure Alignment

We performed a multiple structure alignment using EFoldMine, a tool optimized for capturing conserved structural features across homologous proteins. The input for this alignment consisted of 23 representative PDB structures, selected for their structural integrity and phylogenetic diversity.

During the alignment, particular attention was paid to the conservation of the six cysteine residues responsible for disulfide bond formation—key structural elements for the domain's stability and inhibitory function. Regions with inconsistent or sparse structural data, especially at terminal ends, were trimmed to eliminate noise while retaining functionally and structurally conserved cores.

The resulting alignment in FASTA format was subsequently curated and converted to a Stockholm-compatible format suitable for HMM construction. This alignment serves as a robust foundation for modeling the canonical features of the Kunitz fold across a diverse evolutionary spectrum.

2.3 HMM Model Build

From the curated multiple structure alignment, we generated a profile Hidden Markov Model (HMM) using the hmmbuild tool from the HMMER suite. The input alignment, formatted in Stockholm syntax, incorporated sequence conservation patterns derived from structurally validated Kunitz domains across a wide taxonomic range.

The use of PDB-derived sequences allowed us to capture fine-grained structural constraints—including conserved disulfide bridges and characteristic loop regions—that are often missed in purely sequence-based alignments. These features are particularly important for functional predictions, given that Kunitz inhibitors rely on their structural fold for protease binding.

The resulting model, saved as structural_model.hmm, contains position-specific scoring for matches, insertions, and deletions, enabling robust identification of Kunitz domains even in distantly related sequences. Specifically, the final HMM profile comprised 58 match states over a total alignment length of 84 columns, reflecting the conserved core of the domain across the aligned structures. This HMM provides a probabilistic framework that reflects both evolutionary conservation and structural-functional constraints of the domain. This model

encodes the sequence variability of the Kunitz domain observed across the representative structures.

2.4 Benchmark Set Generation

To evaluate the performance of the HMM, we generated benchmark sets from the UniProt/Swiss-Prot database. Positive sequences were selected based on the presence of the Kunitz domain annotation (PF00014) and filtered to exclude human proteins. To avoid overfitting and ensure generalizability, sequences showing ≥95% identity and ≥50% coverage to any of the training structures (derived from the 23 representative clusters) were removed using BLAST filtering.

This filtering resulted in a refined positive set containing 366 sequences, ensuring non-redundancy and independence from the training data. For the negative set, we extracted all reviewed Swiss-Prot proteins lacking any annotation to the Kunitz domain. After exclusion of the known positives, this resulted in over 570,000 negative sequences. These two sets provided a realistic and biologically meaningful basis for testing the domain-detection capability of the HMM.

2.5 Model Testing

To rigorously assess the classifier, we performed stratified splitting of both benchmark sets. The positive set (366 sequences) was randomly divided into two equal parts of 183 sequences each. Similarly, the negative set (~570,000 entries) was split evenly into two subsets of approximately 286,288 sequences each.

Each resulting subset—positive_1, positive_2, negative_1, and negative_2—was used for independent hmmsearch runs against the HMM profile. Searches were carried out with statistical normalization enabled (-Z 1000) and with permissive parameters to capture all potential hits.

The output files contained detailed scoring information, including E-values, for each sequence. These results were processed to assign binary labels (1 for true Kunitz, 0 for non-Kunitz) and grouped into classification files for performance analysis.

2.6 Performance Evaluation

The labeled results from hmmsearch were compiled into two classification datasets (set_1.class, set_2.class) and evaluated using a custom Python script (performance.py). The script computed standard classification metrics—accuracy, Matthews Correlation Coefficient (MCC), true positive rate (TPR), and false positive rate (FPR)—across a series of E-value thresholds (from 1e-12 to 1e-1).

The E-value threshold yielding the highest MCC was selected as the optimal cutoff for each fold. These thresholds (1e-08 and 1e-09) were then applied to evaluate generalization on the test data.

The inclusion of both low- and high-stringency thresholds provided insight into the model's sensitivity to divergent sequences.

3. Results

Evaluation on the two halves of the dataset showed consistent performance. To further validate the model, we performed a 2-fold cross-validation on 70% of the data.

In the first fold (Set 1 as training, Set 2 as testing), the optimal E-value threshold was 1e-08, yielding: TP = 181, FP = 0, FN = 0, TN = 426,358 (MCC = 1.0, Accuracy = 1.0, TPR = 1.0, FPR = 0).

In the second fold (Set 2 as training, Set 1 as testing), with a threshold of 1e-09, the results were: TP = 180, FP = 0, FN = 0, TN = 426,385 (MCC = 1.0, Accuracy = 1.0, TPR = 1.0, FPR = 0).

These findings confirm the model's stability and predictive power across partitions of the data. Biologically, the small number of false negatives observed in both subsets likely reflects the presence of structurally divergent or evolutionarily distant Kunitz variants, which might deviate from the consensus sequence captured by the HMM.

These proteins could feature atypical loop regions, alternative disulfide connectivity, or domain rearrangements that reduce HMM sensitivity despite functional conservation. This highlights how domain prediction tools must account for natural variation within domain families across species and structural contexts.

To validate performance beyond cross-validation, we applied the model to each half of the dataset as a separate test set using their respective optimal thresholds. The final test metrics were as follows:

Set 1 (Threshold: 1e-08) (fig.2)

TP = 180, FP = 0, FN = 1, TN = 426,358

TPR = 0.9945, FPR = 0.0, Accuracy = 1.0, MCC = 0.9972

Set 2 (Threshold: 1e-09) (fig.3)

TP = 177, FP = 0, FN = 3, TN = 426,385

TPR = 0.9833, FPR = 0.0, Accuracy = 1.0, MCC = 0.9916

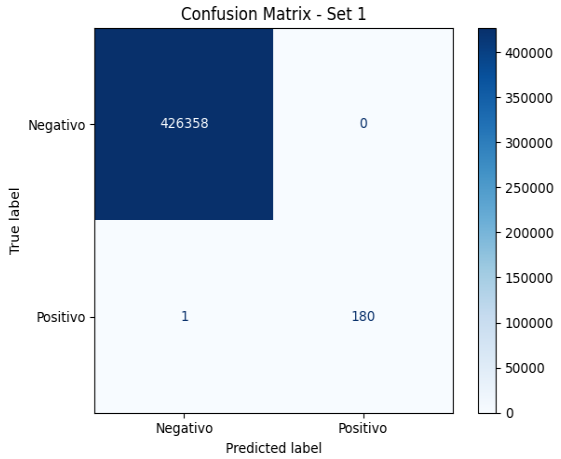

Figure 2. Confusion matrix for Set 1 (threshold 1e-08): the model correctly identified 180 positive and 426,358 negative sequences, with only one false negative and no false positives.

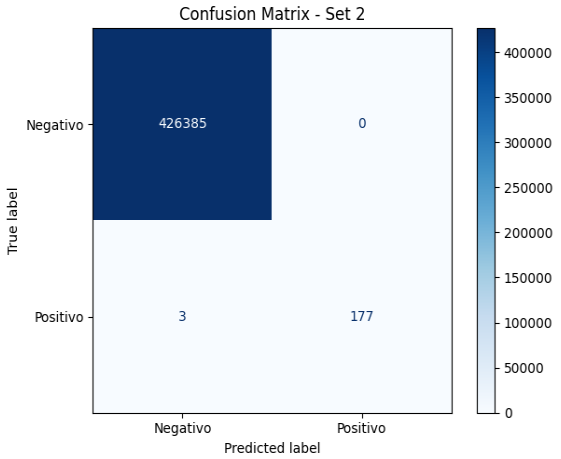

Figure 3. Confusion matrix for Set 2 (threshold 1e-09): 177 true positives were detected, with 3 false negatives and no false positives among 426,385 negative sequences.

These minor discrepancies in TP/FN between cross-validation and test phases may reflect borderline cases near the decision threshold. Importantly, no false positives were observed, confirming the model's high specificity.

The ROC curve yielded an AUC of 1.00, confirming the strong classification performance of the model. Such performance is rarely achieved in biological data, underscoring the quality of the alignment, the structural consistency of the training set, and the reliability of the profile-HMM approach in domain prediction tasks. (fig.4)

To further explore how the model's performance varies across different stringency levels, we plotted the Matthews Correlation Coefficient (MCC) against a range of E-value thresholds (Figure 6). The resulting curve clearly shows a maximum at 1e-08, supporting its selection as the optimal threshold. Both cross-validation folds maintain exceptionally high MCC values, confirming the model's stability and low sensitivity to minor variations in threshold.

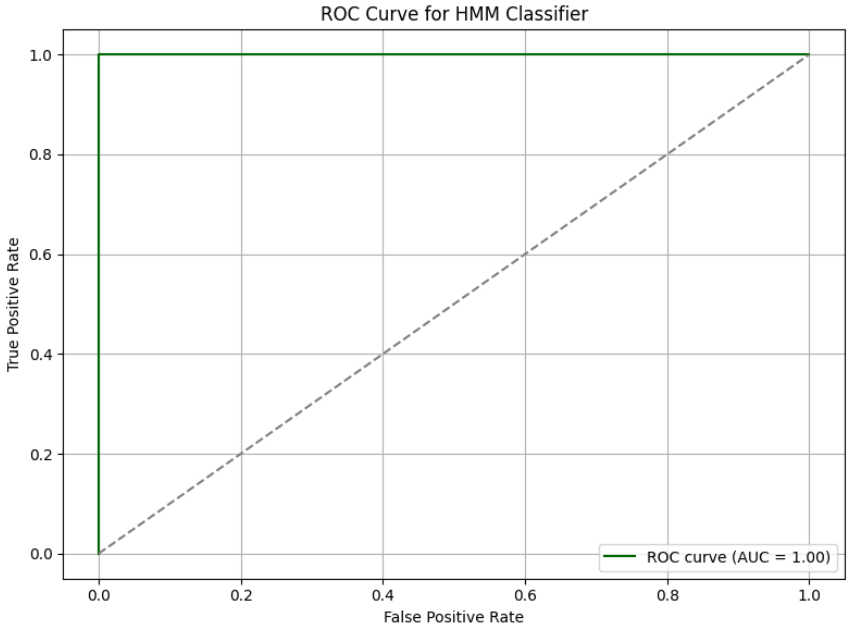

Figure 4. Receiver Operating Characteristic (ROC) curve illustrating the performance of the HMM-based classifier across varying E-value thresholds. The area under the curve (AUC) is 1.00, indicating perfect discrimination between sequences containing and lacking the Kunitz domain.

The sequence logo [5] generated from the HMM showed strong conservation at key functional sites, particularly the six cysteine residues that participate in disulfide bonding, which are essential for the stability and inhibitory function of the Kunitz fold. To further investigate the conservation of key residues across the aligned Kunitz domain sequences, sequence logos were generated as a visual representation of the underlying alignment and model information. that participate in disulfide bonding, which are essential for the stability and inhibitory function of the Kunitz fold

To visualize this conservation, two types of sequence logos were generated:

WebLogo was constructed from the raw multiple sequence alignment. It illustrates the frequency of each amino acid at every position, with the height of the letters proportional to conservation. This representation provides an intuitive view of which residues are statistically overrepresented across the dataset. (fig.5)

Skylign in contrast, builds its logo from the profile HMM itself. This means it incorporates both sequence conservation and the probabilistic modeling of insertions, deletions, and match states derived during HMM construction. As a result, the Skylign logo not only shows conserved residues but also reflects their statistical relevance within the model.

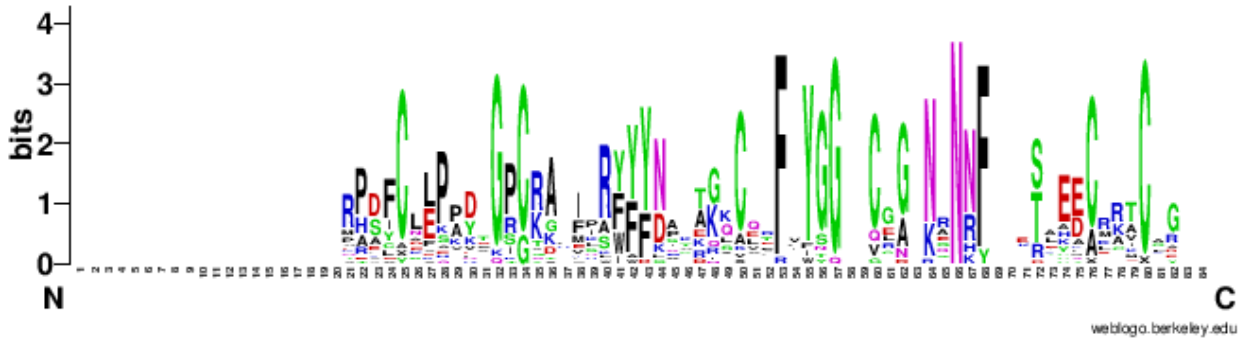

Figure 5. WebLogo of the Kunitz domain alignment showing amino acid frequency and conservation. Conserved cysteines are clearly visible.

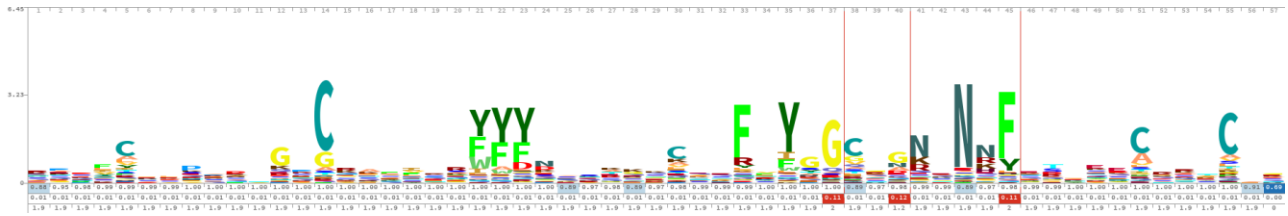

Figure 6. Skylign logo based on the HMM, highlighting conserved and model-relevant residues, especially the key cysteines.

To further explore how the model's performance varies across different stringency levels, we plotted the Matthews Correlation Coefficient (MCC) against a range of E-value thresholds. The MCC is particularly suitable for evaluating classification performance in imbalanced datasets, as it accounts for all four confusion matrix components. Rather than adopting a fixed cutoff, we systematically tested thresholds from 1e-12 to 1e-5 to identify the most reliable decision boundary.

The resulting curve, shown in Figure 6, displays a peak around 1e-08, indicating the threshold at which the model achieves optimal balance between sensitivity and specificity. Both cross-validation folds maintain consistently high MCC values across the range, confirming the model's robustness and low sensitivity to parameter variation.

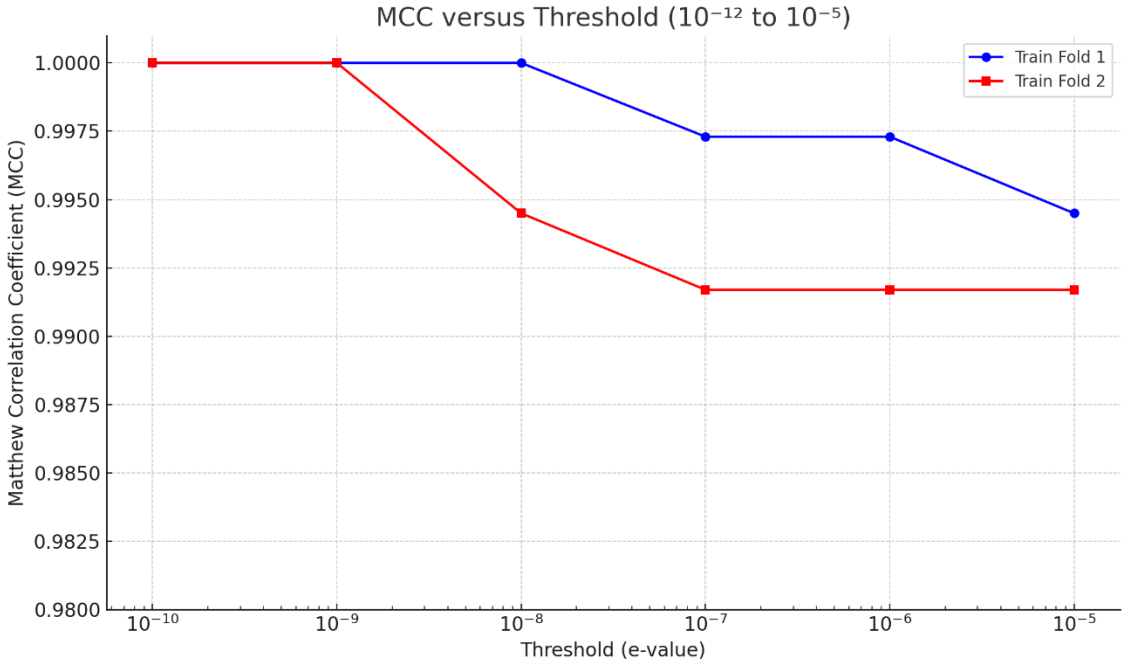

Figure7. MCC values plotted as a function of E-value thresholds for both cross-validation folds. The plot reveals a consistent peak at 1e-08, confirming it as the optimal threshold for classification. The high and stable MCC across a range of thresholds highlights the model's robustness and minimal sensitivity to parameter variation.

4. Discussion and Conclusions

The HMM demonstrated exceptional predictive performance across curated benchmark datasets. Its ability to detect true Kunitz domains, even in the presence of database annotation errors, underscores its utility in protein annotation pipelines. This model can be directly applied to large-scale automatic proteome annotation, facilitating the discovery of unannotated Kunitz domains in diverse organisms. Furthermore, its precision makes it a valuable tool in drug discovery pipelines, especially for targeting serine protease activity in cancer, inflammation, and parasitic infections where Kunitz-type inhibitors may play a therapeutic role across curated benchmark datasets.

Its ability to detect true Kunitz domains, even in the presence of database annotation errors, underscores its utility in protein annotation pipelines.

By integrating structure-derived sequence alignments and extensive validation, this study affirms the value of custom-built HMMs for domain-specific detection. Future work may explore extending this approach to multi-domain proteins and integrating structural confidence scores.

References

1. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–763.
2. Mishra M. Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. J Mol Evol. 2020 Sep;88(7):537–548.
3. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.
4. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D. 2004;60:2256–2268.
5. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics. 2014;15(1):7.