Hidden Markov Model profile for the Kunitz-type protease inhibitor domain

Rossi Anna

Project report for Laboratory of Bioinformatics 1 (MSc in Bioinformatics University of Bologna)

Abstract
Protease inhibitors play critical roles in regulating proteolytic activity in biological systems, and the Kunitz domain constitutes one of the most extensively studied inhibitory clusters. A computational pipeline was implemented to construct and calibrate a profile Hidden Markov Model (HMM) for the detection of the Kunitz-type protease inhibitor domain (Pfam ID: PF00014), exemplified by the bovine pancreatic trypsin inhibitor (BPTI). A multiple sequence alignment (MSA) of canonical Kunitz-type proteins was generated via structural superposition using PDBeFOLD, and the resulting alignment was employed to train the HMM. Model performance was assessed on positive (Kunitz-type) and negative (non–Kunitz-type) datasets using a two-fold cross-validation scheme. Sensitivity was demonstrated by the majority of true positives scoring E-values well below the standard threshold (1×10^{-3}). A comprehensive suite of statistical metrics, including confusion matrix, overall accuracy (Q_2), recall, precision, Matthews Correlation Coefficient (MCC), and area under the ROC curve (AUC), was calculated to confirm the model's discriminative power and reliability.

Supplementary information: Supplementary data are available on the following GitHub link

1 Introduction

The Kunitz domain is a compact (~50–60 amino acid) α+β fold characterized by three conserved disulfide bonds which confer exceptional stability and enable potent inhibition of serine proteases. Prototypical examples include bovine pancreatic trypsin inhibitor (BPTI), tick anticoagulant peptide (TAP), and domains within human amyloid precursor protein (APP) and tissue factor pathway inhibitor (TFPI) [1] High-resolution structures, such as PDB 3TGI of the rat trypsin–BPTI complex (***Figure 1***), reveal a twisted, two-stranded antiparallel β-sheet packed against a short α-helix, with six cysteines forming three disulfide bridges that lock the fold. (***Figure 2)***

Kunitz-domain protease inhibitors regulate critical physiological processes by blocking aberrant proteolysis; for instance, TFPI modulates coagulation by inhibiting factor Xa and the TF–VIIa complex, while BPTI analogs protect tissues from uncontrolled trypsin activity. [2] Plasma inter-α-trypsin inhibitor (ITI), a multi-chain Kunitz-type inhibitor, plays key roles in extracellular matrix stabilization and inflammation by covalently linking glycosaminoglycans to proteins. [3] Kunitz domains are also exploited as scaffolds for engineered protease inhibitors in therapeutic design, owing to their stability, small size, and high specificity. [4]

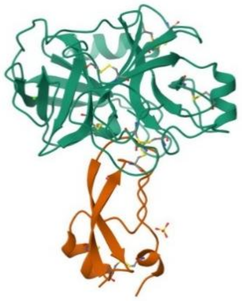

Fig.1 PDB structure of 3TGI: Wild-type rat anionic trypsin complexed with bovine pancreatic trypsin inhibitor (BPTI). The protein is composed by two chains: chain-A(green) represent the trypsin part and chain-B(orange) represents the Kunitz Domain

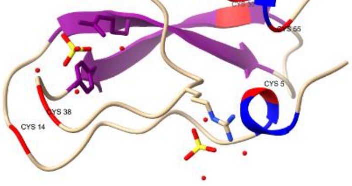

Fig.2 Detailed visualization of the Kunitz domain structure with ChimeraX In purple are highlighted the two-stranded antiparallel β-sheet and in blue the α-helix. In red are showed the 6 fundamentals cysteine residues of the domain(CYS).

Hidden Markov Models (HMMs) are statist representations designed to capture the position-specific conservation and variability observed in families of biological sequences. An HMM pattern a sequence family by combining match, insertion, and deletion states arranged linearly to reflect a multiple sequence alignment (MSA). Each match state emits residues according to the observed frequency at that alignment column, while insertion and deletion states model gaps explicitly, allowing the HMM to naturally accommodate indels. [5] The probabilistic framework of HMMs enables calculation of the likelihood that a given query sequence belongs to the modeled family, yielding both bit-scores and E-values for rigorous statistical interpretation of matches.

Why are HMM models ideal for Kunitz-Domain detection?

1. In the Kunitz domain, where cysteines forming disulfide bridges and reactive-site loops are critically conserved, profile HMMs ensure that substitutions at these key positions are heavily penalized, while more variable regions are modeled with greater flexibility

2. Profile HMMs excel at detecting distantly related sequences with low overall identity, critical for discovering novel Kunitz inhibitors in under-explored taxa. Empirical benchmarks demonstrate that HMMER-based profile searches outperform BLAST-type algorithms, achieving comparable speed to BLAST while retaining superior sensitivity through probabilistic inference [6]

By constructing a tailored profile HMM from a curated MSA of experimentally validated Kunitz domains, the aim is to capture both the conserved structural scaffold (e.g., the six-cysteine framework) and the allowed sequence diversity. The resulting model will be deployed with HMMER's tools to scan large protein datasets, enabling the discovery of novel Kunitz-type protease inhibitors. This approach leverages the statistical rigor and sensitivity of profile HMMs to expand our functional annotation of protease inhibitor repertoires and to identify candidates for therapeutic and biotechnological applications.

2 Methods

2.1 Training set preparation

The HMM model initially was initially trained on a training set of well-known Kunits Domain proteins, easily obtained from *PDB (Protein Data Bank)* using the advance search with these criteria:

- Data Collection Resolution <= 3.5
- Identifier - Pfam Protein Family: PF00014
- Polymer Entity Sequence Length <= 80 and >= 45

These criteria were selected to ensure high-quality structural data (resolution ≤ 3.5 Å) and to focus specifically on proteins annotated as part of the Kunitz domain family (*Pfam: PF00014*). The sequence length range reflects the typical size of Kunitz domains, helping to exclude outliers and ensure the training set represents the canonical structure of this protein family.

The quality and consistency of the input sequences were performed using these two processing steps:

a. Redundancy reduction with CD-HIT

Using *CD-HIT (v.4.8.1)* [7] the initial set of 160 protein sequences was clustered based on sequence identity, removing the redundancy. CD-HIT efficiently groups sequences sharing high similarity, retaining only representative sequences from each cluster. This step reduces computational burden and prevents bias in the multiple sequence alignment (MSA) due to overrepresented sequences. The output express 25 representative sequences, one for each cluster.

b. Manual filtering of outliers

One sequence was removed from the CD-HIT output due to excessive length, which could distort the MSA by introducing large gaps or misalignments, given that the typical Kunitz domain ranges between 45 and 80 residues.

2.2 Multiple Sequence Alignment (MSA)

The filtered set was aligned using *PDBeFOLD* [8] to account for conserved features of the Kunitz domain. This tool allows the comparison of protein 3D structures by aligning their secondary structure elements to identify structural similarities. Prior to finalizing the MSA, was removed one additional sequence that showed a root mean square deviation (RMSD) greater than 1 Å compared to the consensus structure.

RMSD (Root Mean Square Deviation) is a measure of the average distance between corresponding atoms of two superimposed protein structures. A high RMSD may indicate a structural divergence that could introduce noise into the alignment and compromise the specificity of the resulting HMM.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Fig.3 RMSD formula: where δ_i is the distance between atom i and either a reference structure or the mean position of the N equivalent atoms. This is often calculated for the backbone heavy atoms C, N, O, and C_α or sometimes just the C_α atoms

These preprocessing steps ensure that the final MSA used to train the HMM is composed of non-redundant, high-quality, and structurally consistent sequences, improving the accuracy and biological relevance of the model.

2.3 HMM model construction

Using the *hmmbuild* tool from HMMER (v.3.4), a Hidden Markov Model (HMM) profile for the Kunitz domain was built using the multiple sequence alignment (MSA) previously generated. The resulting HMM includes 58 match states, corresponding to the most conserved and well-aligned positions across the sequences in the MSA. Out of the total 84 alignment columns, these match states represent the core regions of the domain where evolutionary conservation is strongest. To visualize the conservation pattern across the domain, a sequence logo was generated using Skylign, highlighting amino acid preferences at each match state.

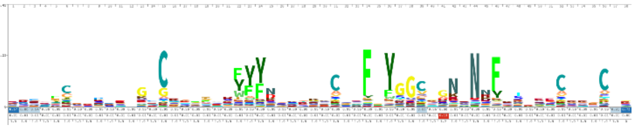

Fig.4 SL-HMM (Skylign) It shows clearly the 6 conserved cysteine involved in the three disulfide bridges that lock the fold

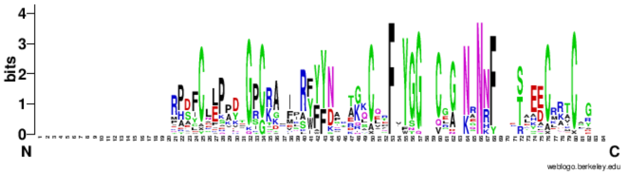

Fig.5 SL-MMA (Weblogo) This logo represents the full alignment, including all columns, and highlights conserved residues across the dataset before HMM construction.

To visually compare sequence conservation before and after HMM construction, two sequence logos were generated: one from the multiple sequence alignment (**Fig.5**) using *WebLogo* (https://weblogo.berkeley.edu/logo.cgi), and one from the HMM profile (**Fig.4)** using *Skylign* (https://skylign.org/). As expected, the two logos appear similar, reflecting the consistency of the input alignment and the resulting model. However, minor differences are present. The SL-HMM logo contains fewer positions, as *hmmbuild* filters out poorly aligned columns and focuses only on conserved "match states." Additionally, some residues in SL-HMM appear slightly more prominent, indicating that the HMM assigned them greater weight due to higher conservation or lower gap frequency. The presence of the six conserved cysteines, characteristic of the Kunitz domain, in both logos confirms the quality of the alignment and the effectiveness of the model.

2.4 Test sets preparation

To evaluate the performance of the constructed HMM model, two distinct test sets were prepared: one positive and one negative. The positive test set was obtained from *UniProt* using the *Advanced Search* tool, by selecting all reviewed (*Swiss-Prot*) proteins annotated with the *Pfam* domain PF00014, which corresponds to the Kunitz domain. The negative test set was built by extracting all reviewed Swiss-Prot proteins not containing the Kunitz domain. This strategy leverages the high curation quality of Swiss-Prot entries, ensuring accurate domain annotation and minimizing false positives or misannotations.

To further ensure the robustness of the evaluation, the positive test set was filtered for redundancy using *BLAST* (*Basic Local Alignment Search Tool*). Specifically, sequences exhibiting ≥ 95% sequence identity and ≥ 50% alignment coverage either to each other or to sequences from the training set were removed. This filtering step was applied to avoid inflated performance metrics due to excessive sequence similarity, and to ensure the model is evaluated on non-redundant, non-overlapping data.
Thanks to this process we found 29 redundancy proteins between the training set and the positive test set.

Table 1. Positive and negative test sets

Test set	Number of proteins
Positive test set (PF00014)	397 – 29 = 369
Negative test set	572833

To prepare for cross-validation and ensure a balanced evaluation, both the datasets were randomly shuffled and then split into two equal subsets. This strategy enables the construction of independent training and testing partitions while preserving class balance. Randomization reduces potential biases due to ordering or grouping of similar sequences and contributes to the robustness and generalizability of the model's performance evaluation.

2.5 HMM model evaluation

To assess the predictive capacity of the constructed HMM model, *hmmsearch* was performed on the four distinct test sets obtained by randomly shuffling and splitting the positive and negative test sequences into two halves each. The goal of this split is to ensure independent evaluation subsets, reducing bias and overfitting during threshold tuning and performance analysis. After running *hmmsearch*, the results were parsed into classification tables containing sequence IDs, true labels, scores, and E-values.
HMMER filters by default out matches with E-values >10.0; these were manually recovered and assigned an E-value of 10.0 to avoid bias.

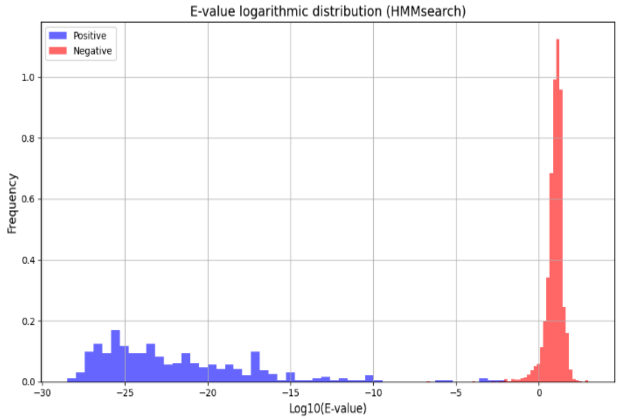

Fig.6 Distribution of e-values in log scale (log10) for positive and negative sets. The positive set (blue) and the negative set (red) are normalized to density, allowing for a clear comparison despite the imbalance in their sizes. This visualization helps to assess the performance of the HMM model by comparing the shape of the distributions, with lower e-values typically indicating the presence of the domain. For the graphic was used the four results of the *hmmsearch* performed on the four obtained sets (two positive and two negative).

2.6 Two-fold cross-validation

Following the initial classification, to further evaluate the model's performance was implemented a *2-fold cross-validation scheme*. In each run, one fold served as the training subset for potential threshold optimization, while the other was used for independent testing. This allowed us to assess the stability and generalizability of the HMM model across different portions of the data. Performance metrics such as accuracy, precision, recall, confusion matrix and MCC were performed. This strategy, often used in machine learning and biological sequence classification [9], ensures that performance evaluation is not overly reliant on a single train/test split.

3 Results

The model achieved strong discriminatory power while balancing sensitivity and specificity. During 2-fold cross-validation, multiple E-value cutoffs were tested, assessing MCC alongside accuracy, confusion matrices, precision, and recall (***Figure*** *7*), and identified 1e-3 for Set 1 and 1e-2 for Set 2. To err on the side of conservatism, the 1e-3 threshold was used for extracting classification errors: sequences with label 1 and E-value ≥ 1e-3 were deemed false negatives, and those with label 0 and E-value < 1e-3 false positives.

At this cutoff, two true Kunitz proteins (*Q8WPG5*, E-value 0.002; *D3GGZ8*, E-value 0.0096) fell above the threshold and were misclassified. UniProt annotation, however, confirms a Kunitz domain in both.

In **Table 2**, all statistical results of the model's performance are reported for the threshold corresponding to the best E-value. ***Figure 8*** displays the corresponding confusion matrices, including the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Additionally, a receiver operating characteristic (ROC) curve was generated to assess the model's overall discriminative performance by plotting the true positive rate (TPR) against the false positive rate (FPR) across a range of E-value thresholds. The resulting area under the curve (AUC = 1.00) provides a quantitative measure of the model's classification capability (***Figure 9***).

The ROC curve was computed by combining the two evaluation sets obtained from 2-fold cross-validation and calculating TPR and FPR over multiple threshold values. As expected, the curve reaches the upper-right corner (TPR = 1.0, FPR = 1.0) at lower stringency thresholds, where the model predicts almost all samples as positive. However, this behaviour does not reflect the model's performance under optimal classification conditions. At the best-performing thresholds identified for each fold, the confusion matrices (Figure 8) indicate excellent specificity and sensitivity, with an MCC of 1.00 and only two false negatives across the sets. Therefore, while the ROC curve offers a global perspective on model behavior across thresholds, the fold-specific performance metrics confirm its practical reliability and robustness.

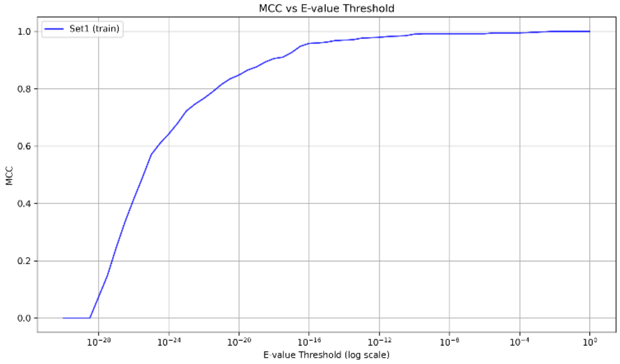

Fig.7 MCC vs E-value threshold Multiple E-value threshold was tested for obtain the best possible MCC value (1.00), showed at a cutoff between 10e-6 and 10e0

Table 2. Performance results The following metrics are reported: Accuracy (Q2), Matthews Correlation Coefficient (MCC), Recall (True Positive Rate or TPR), and Precision (Positive Predictive Value or PPV).

	Best threshold	Accuracy (Q2)	MCC	Recall (TPR)	Precision (PPV)
Set 1	1e-3	1.0000	0.9945	0.9891	1.0000
Set 2	1e-2	1.0000	1.0000	1.0000	1.0000

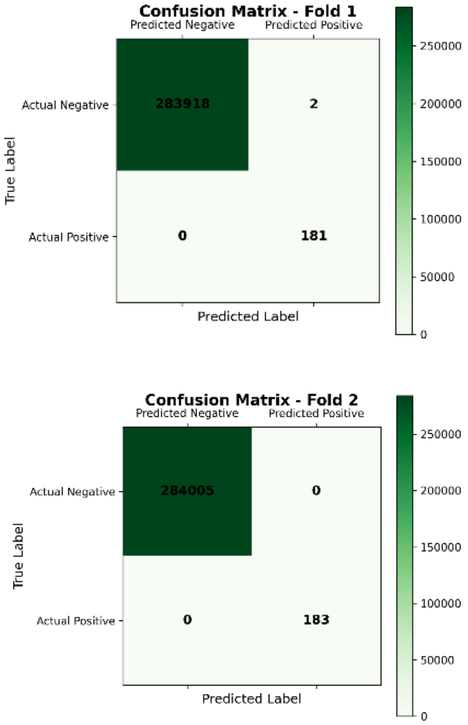

Fig.8 Confusion matrices In the Set 1 confusion matrix, two positive instances were misclassified as negative (FN = 2), indicating that a small number of true signals failed to be detected at the chosen decision threshold. This contributes to a slight reduction in sensitivity for Set 1, although specificity remains high. The balance between these error types is further quantified by the ROC curve (Figure 9), whose AUC reflects the overall trade-off between true positive and false positive rates across thresholds.

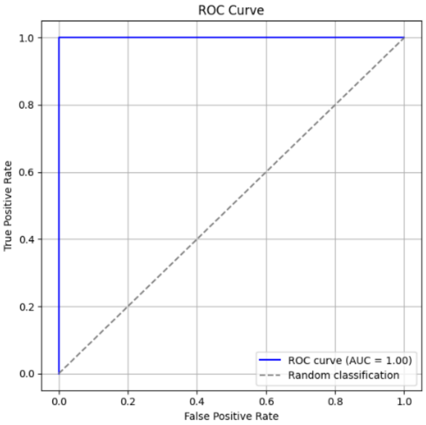

Fig.8 ROC curve The model achieves an AUC of 1.00, indicating perfect classification performance across varying E-value thresholds.

4 Discussion and conclusion

The results obtained with the HMM model for Kunitz domain identification are overall highly promising. In both folds of the cross-validation, perfect accuracy (Q2 = 1.0000) was achieved, with an average Matthews Correlation Coefficient (MCC) of 0.9973 and an area under the ROC curve (AUC) of 1.00, confirming the model's exceptional discriminative power at the optimal thresholds (1×10^{-3} and 1×10^{-2}). The only deviation of the model was represented by two false negatives (*Q8WPG5* and *D3GGZ8*), proteins annotated in UniProt as containing the Kunitz domain but not recognized by the model at the chosen threshold. Thissuggest that these cases may involve family variants with sequences too divergent from the training set, or partially annotated or atypical structural arrangements of the conserved regions.

A targeted analysis of these two proteins, such as multiple sequence alignment against the HMM profile, examination of low-complexity regions, or investigation of indels in key domain sites, could clarify whether they represent model limitations or annotation discrepancies Going forward, adding these tough cases into our training data and fine-tuning the HMM settings should boost our ability to catch all true Kunitz domains without losing any of the model's current accuracy.

References

[1] InterPro Entry IPR002223 – Kunitz/Bovine pancreatic trypsin inhibitor domain. European Bioinformatics Institute (EMBL-EBI).
Available at: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002223/

[2] Nixon AE, Wood CR. Engineered protein inhibitors of proteases. Curr Opin Drug Discov Devel. 2006 Mar;9(2):261-8. PMID: 16566296.

[3] Salier JP. Inter-alpha-trypsin inhibitor: emergence of a family within the Kunitz-type protease inhibitor superfamily. Trends Biochem Sci. 1990 Nov;15(11):435-9. doi: 10.1016/0968-0004(90)90282-g. PMID: 1703675.

[4] InterPro Entry IPR002223 – Kunitz/Bovine pancreatic trypsin inhibitor domain, Pathways section. European Bioinformatics Institute (EMBL-EBI). Available at: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002223/pathways/#table

[5] Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep;10(6):402-15. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.

[6] Zhang Y, Sun Y. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics. 2011 May 24;12:198. doi: 10.1186/1471-2105-12-198. PMID: 21609463; PMCID: PMC3115854.

[7] Li, W., & Godzik, A. (2006). *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*. Bioinformatics, 22(13), 1658–1659.

[8] Krissinel, E., & Henrick, K. (2004). **Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions**. *Acta Crystallographica Section D: Biological Crystallography*, 60(12), 2256–2268.

[9] Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 (IJCAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137–1143.