Bioinformatics, 2025, 05–08
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 08th May 2025

Bioinformatics

Building a Profile Hidden Markov Model for the Kunitz-Type Protease Inhibitor Domain

Marco Cuscunà[1]

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

Abstract

Motivation: Protein domains are critical in determining protein structure and function. The Kunitz-type protease inhibitor (KTPI) domain is an evolutionarily conserved motif involved in protease inhibition. This study focuses on building and evaluating profile Hidden Markov Models (HMMs) for the identification of the Kunitz domain in proteins, comparing sequence-based and structure-based approaches.

Results: Two profile HMMs were constructed using the HMMER library: one based on a multiple sequence alignment (MSA) using MUSCLE, and another on a structure-based alignment (MStA) generated with PDBeFold. Both models were evaluated via 2-fold cross-validation using curated UniProt datasets. In the Pfam-based evaluation, both models showed excellent performance (Avg MCC ≈ 0.9945), with the structure-based model achieving slightly better results compared to sequence one, under single-domain scoring (MCC = 0.993182 vs 0.993180). A further evaluation was carried out to address a key limitation of the initial analysis: the reliance on Pfam annotations (PF00014) for defining the positive set. It became evident the protein classified as false positive under single-domain scoring by both models was in fact a true Kunitz domain lacking Pfam annotations due to cross-reference inconsistencies. To better capture these cases, an expanded analysis was performed using the InterPro ID IPR036880, which also defines the Kunitz domain. Unexpectly, average MCC scores decreased slightly (e.g., 0.9843 for the structure-based model), reflecting a higher number of apparent false positives. However, manual inspection confirmed that all of these were genuine Kunitz proteins, previously missed due to incomplete annotation. This drop in MCC does not indicate a loss in performance; rather, it highlights the models' ability—especially the single-domain structure-based HMM—to uncover biologically relevant instances beyond the limits of current annotation databases. These findings reinforce the importance of incorporating structural information into HMM-based protein domain detection and highlight the value of model-driven inference for uncovering domain occurrences that are missed by traditional annotation pipelines.

Availability: All code, data, and supplementary materials are available at: https://github.com/Markus2409/MSc_Bioinformatics_HMM_KunitzDomain

Contact: marco.cuscuna@studio.unibo.it

1. Introduction

Multiple sequence alignment is a crucial step in assessing homology between proteins, identifying functional similarities, and tracing their evolutionary history. However, sequence-based alignments can struggle to capture distant relationships, particularly when sequence identity is low or when the homology of full-length genes is uncertain. In such cases, structural information becomes especially valuable. That's because structural alignments are generally more reliable because protein structures are more conserved than their sequences (Carpentier et al., 2019 [15]). This structural conservation allows for more accurate identification of homologous residues and domains, even when the corresponding sequences have diverged significantly. Understanding the intricate interplay between a protein's three-dimensional structure and its biological function is therefore essential. The spatial arrangement of a protein's atoms determines not only its stability but also its specific interactions with other molecules, and ultimately its role in biological systems. This report focuses specifically on the Kunitz-type protease inhibitor domain.

1.1 Kunitz-type protease inhibitor domain

Kunitz-type protease inhibitor domains are found in a wide variety of proteins and are well known for their potent anticoagulant and protease inhibitory activities. These domains are frequently present in toxins and venoms from a broad range of organisms including wasps, spiders, scorpions, and snakes, and have been extensively studied for their ability to inhibit serine proteases and reduce blood coagulation (Fratini et al., 2022 [16]).

This anticoagulant property makes Kunitz domains of great clinical interest, particularly in the development of therapeutic agents designed to reduce bleeding during complex surgical procedures, such as heart and liver surgeries.

Remarkably, Kunitz domains are also found in non-toxic, human proteins, such as the amyloid precursor protein (APP). Certain isoforms of APP contain KPI (Kunitz Protease Inhibitor) domains, which have been linked to the worsening of mitochondrial dysfunction—a key hallmark of Alzheimer's disease (Chua et al., 2013 [14]).

The Bovine Pancreatic Trypsin Inhibitor (BPTI) is the prototypical member of this family and was the first Kunitz-type inhibitor described (Kunitz & Northrop, 1936[12]). BPTI exhibits broad specificity, inhibiting trypsin, chymotrypsin, and elastase-like serine proteases (Ascenzi et al., 2003 [1]). This family of domains is also well-defined in domain annotation databases, such as Pfam, where it is indexed as PF00014, and is widely used as a reference for protein classification and dataset construction in computational biology.

Kunitz domains may occur as single units (e.g., in BPTI), or may be repeated multiple times within the same polypeptide chain, forming multi-domain inhibitors. For instance, the Tissue Factor Pathway Inhibitor (TFPI) contains three Kunitz domains, while the parasitic hookworm *Ancylostoma caninum* expresses a Kunitz-type inhibitor with 12 tandem domains, allowing independent interactions with multiple proteases via distinct reactive sites. Structurally, the Kunitz motif consists of a peptide chain of approximately 60 amino acids, stabilized by three conserved disulfide bridges.

The Kunitz domain adopts a disulfide-rich α+β fold, which is stabilized by three highly conserved disulfide bonds, typically formed between the following cysteine pairs: C1–C6, C2–C4, and C3–C5 (Figure 1). The C1–C6 and C3–C5 bridges are essential for maintaining the native conformation (Creighton, 1975 [11]), while the C2–C4 bond stabilizes the protease-binding loop and enhances inhibitory function (Laskowski & Kato, 1980 [10])

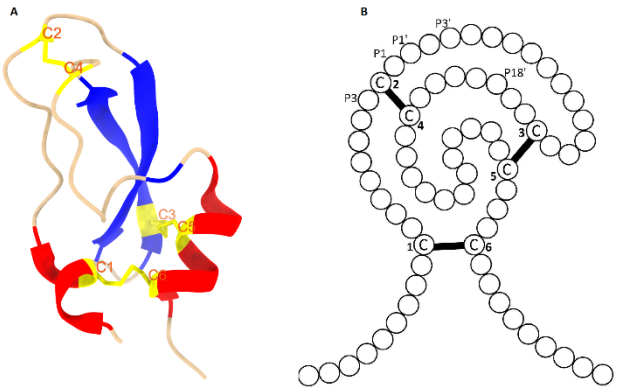


Figure 1 Predicted folding structure of a single-domain Kunitz-type protease inhibitor. (A) 3D representation of the domain from BPTI (PDB: 1BHC, chain A), rendered using UCSF ChimeraX. The structure shows the canonical α+β fold, with α-helices (red), β-strands (blue). The three highly conserved disulfide bridges (yellow)—C1–C6, C2–C4, and C3–C5—stabilize the compact conformation. (B) Schematic 2D diagram highlighting the position and pairing of the six cysteine residues, forming the three disulfide bonds. The loop between positions P_3 and P'_3 represents the protease binding loop, crucial for inhibitory activity (modified from Chand et al., 2004 [13]).

1.2 Hidden Markov Models

Hidden Markov Models (HMMs) are powerful statistical tools widely used in computational biology to model and analyze biological sequences such as proteins and nucleic acids.

An HMM consists of two stochastic processes (Figure 2):

- A hidden state process, which determines the current context (e.g., domain, motif, or structural feature),
- And a visible process, which emits observable symbols (such as residues) based on the current hidden state.

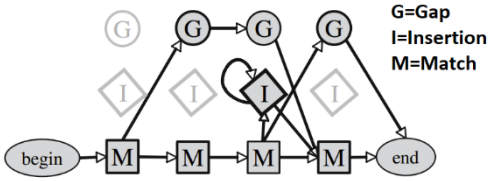
The Model, in this way, describes a sequence of observable events assuming that the probability distributions of these visible symbols depend on the underlying, unobservable (hidden) process (Yoon, 2009 [9]).

The transition probabilities define how the system moves between hidden states, forming the model's topology, while the emission probabilities describe the likelihood of observing specific residues or symbols in each state. Together, these define a probabilistic framework that can capture both conservation and variability across related biological sequences.

In protein domain analysis, HMMs are particularly effective because they can model position-specific features, including conserved residues, insertion-deletion patterns and evolutionary variability.

This is achieved by training an HMM on a multiple sequence/structure alignment of known domain instances, yielding what is called a profile HMM. This profile provides a formal probabilistic model that is robust and flexible in detecting remote homologs, even in the presence of noise or low sequence identity.

For this reason, HMMs have become a standard tool in sequence annotation, domain detection, and protein family classification, and are used extensively in databases such as Pfam, SMART, and InterPro.



Visible States	Hidden States
ADG••K	MMMM
A••••K	MGGM
ADG•••	MMMG
ADGHIK	MMMIIM

Figure 2 Each hidden state (M = Match, I = Insertion, G = Gap) generates observable symbols corresponding to amino acid residues (e.g., A, D, G, K). The top part illustrates the model's architecture, with possible state transitions. The table below highlights the distinction between visible states (aligned residues, including gaps "•") and hidden states, which represent the evolutionary nature of each position: conserved matches (M), insertions (I), and gaps (G). The HMM estimates transition and emission probabilities, enabling the optimal identification of functionally relevant regions in protein domains (modified from Bystroff & Krogh, 2008 [8]).

1.3 Structural-based vs Sequence-based Alignment

Multiple sequence alignments (MSAs) play a fundamental role in bioinformatics analyses, particularly in understanding functional relationships and inferring the evolutionary history of biological macromolecules. MSAs compare the primary sequences of proteins or nucleic acids, arranging them in a matrix where each column ideally represents homologous residues. These columns are assumed to correspond to structurally superimposable or functionally equivalent positions in the aligned sequences (Edgar & Batzoglou, 2006 [7]).

To support these analyses, several tools have been developed to perform efficient and accurate MSAs, such as MUSCLE (Edgar, 2004 [6]). However, it is well established that protein sequences undergo evolutionary changes over time, including point mutations, insertions, deletions, and recombination events, which can significantly alter the primary sequence while preserving structural or functional characteristics. As a result, sequence-based alignments may lose reliability when sequence identity drops below detectable thresholds.

For this reason, the use of multiple structural alignment (MStA) methods has become increasingly important. Unlike sequence alignments, these methods compare the three-dimensional conformations of proteins, offering insights that are not apparent from sequence data alone (Carpentier et

al., 2019 [15]). Structural alignments aim to maximize the spatial correspondence between protein backbones or specific residues by optimizing scoring functions such as RMSD (root-mean-square deviation), Q-score, or TM-score.

This approach is particularly valuable in cases of remote homology, where two proteins share structural similarity despite having low sequence similarity due to deep evolutionary divergence. In such scenarios, structure-based alignments provide a more robust and accurate basis for functional annotation, domain identification, and comparative modeling.

1.4 Aims

The aim of this analysis is to construct a profile Hidden Markov Model (HMM) for the Kunitz-type protease inhibitor domain, using as input the most representative sequences annotated with Pfam PF00014, extracted from the Protein Data Bank (PDB). The model is subsequently tested on a dataset composed of all Kunitz-domain proteins available in SwissProt, excluding those used for model construction, in order to identify the optimal E-value threshold that minimizes false positives and false negatives, and to evaluate the predictive performance of the model.

Model validation was performed through a 2-fold cross-validation strategy, in which the dataset is split into two subsets: one used for training and the other for validation, then swapped in a second iteration to ensure a balanced and robust performance estimate. In addition to this cross-validation, a third test was conducted on the entire combined dataset (overall) to assess the model's global performance.

For each of these three evaluation settings (SET1, SET2, and OVERALL), the performance was assessed using two different classification modes:

- based on the E-value computed from the full protein sequence ("full sequence");
- based on the E-value only for the single best-scoring domain found in the sequence ("best 1 domain").

The entire procedure was repeated for both an HMM constructed from a multiple sequence alignment (MSA) and an HMM based on a multiple structural alignment (MStA). This dual approach allowed a direct comparison of the two alignment strategies, to evaluate which one provides better performance for the construction of HMMs targeting the Kunitz domain.

2. Methods

The complete documentation, including scripts, datasets, and results, is publicly available and accessible via the GitHub repository linked in the abstract.

2.1 Data Preparation

To construct and evaluate the Hidden Markov Models (HMMs), a curated set of Kunitz-type domain sequences was retrieved from the Protein Data Bank (PDB) using Pfam ID PF00014 as a reference for domain annotation. The selection was carried out using the following advanced query parameters:

- Data Collection Resolution ≤ 3.5 Å
- Pfam Annotation = PF00014
- Polymer Entity Sequence Length between 45 and 80 residues

This query was designed to obtain high-quality crystal structures that contain a single, well-defined Kunitz domain. The retrieved entries were saved in the file rcsb_pdb_custom_report_20250410062557.csv.

To reduce redundancy and obtain a representative set of sequences, the entries were clustered using CD-HIT with a 90% sequence identity threshold, a widely used tool in bioinformatics for sequence redundancy reduction (Fu et al., 2012 [5]). This process yielded 25 clusters, from which the most representative sequence of each cluster was selected.These representative sequences were then saved into the file pdb_kunitz_rp.fasta, which served as the reference dataset for downstream analyses. A manual check was performed to assess consistency in sequence lengths among the selected representatives. The goal was to ensure that no sequence would introduce bias in the alignment due to unusually long insertions or extensions. One outlier sequence, 2ODY:E, was found to be significantly longer than the others and was therefore excluded from the dataset.

The resulting dataset was used to extract a list of compatible PDB IDs (tmp_pdb_efold_ids.txt) formatted for use with PDBeFold. This set served as the reference for both the multiple sequence alignment (MSA) and the structural alignment (MStA) procedures. The entire pipeline for data collection and preprocessing is encapsulated in the script script_recover_representative_kunitz.sh.

Additional datasets were downloaded from UniProt/Swiss-Prot, including:

- all_kunitz.fasta: all Kunitz proteins filtered by Pfam annotation PF00014.
- uniprot_sprot.fasta: the full Swiss-Prot database, used as the background for negative dataset construction.

2.2 Multiple structure alignment

A multiple structural alignment was performed using PDBeFold, based on the list of representative PDB IDs previously extracted. Upon inspection of the results, one structure (5JBT:Y) showed a significantly high RMSD value of 2.9180, compared to the average RMSD across the other domains (typically in the range of 0.4–0.5 Å). Due to its outlier behavior, this domain was excluded from the dataset.

The alignment was then recomputed and saved in pdb_kunitz_rp_strali.fasta, resulting in the final multiple structure alignment, which served as the basis for subsequent model construction.

2.3 Multiple sequences alignment

To ensure consistency between the structural and sequence alignments, the same representative sequences used for the multiple structural alignment were reused for the sequence-based analysis.

These sequences were re-extracted from pdb_kunitz_rp.fasta using the same set of IDs previously used for the PDBeFold alignment—already cleaned of potential outliers—and saved in pdb_kunitz_rp_clean.fasta.

The cleaned dataset was then aligned using MUSCLE (Edgar, 2004 [6]), a widely used algorithm for multiple sequence alignment, and saved in pdb_kunitz_rp_seqali.fasta

2.4 HMMs building a Test set generation

Profile Hidden Markov Models (HMMs) were built using the HMMER suite, a widely used collection of tools for sequence analysis based on probabilistic models. HMMER enables sensitive detection of homologous sequences by modeling position-specific residue frequencies, insertions, deletions, and alignment uncertainties (Eddy, 2010 [2]). In this study, HMMs were generated based on both the multiple sequence alignment and the multiple structural alignment previously obtained. The resulting models were saved as pdb_kunitz_rp_seqali.hmm and pdb_kunitz_rp_strali.hmm, respectively.

The same approach was used for generating testing sets for both models (MSA and MStA). To prevent overlap between training and testing data, a BLAST search was performed to identify sequences in the Swiss-Prot Kunitz dataset (all_kunitz.fasta) with high similarity (≥95% identity) to the training sequences in pdb_kunitz_rp_clean.fasta. All matching entries

were excluded. The remaining sequences were randomly divided into two positive sets (pos_1 and pos_2), which served as the positive examples for SET 1 and SET 2. Their sequences were extracted from the Swiss-Prot database using the script get_seq.py.

For the negative sets, all Kunitz-related IDs (included the Kunitz sequences used to build HMMs) were removed from Swiss-Prot to obtain a background of non-Kunitz proteins. These were also randomized and split into two groups (neg_1 and neg_2), serving as the negative examples for the respective sets.

Each of the four sets (positive and negative for SET 1 and SET 2) was scanned using the appropriate HMM. The resulting output files contained two statistical scores for each sequence: the E-value for the full protein and the E-value of the best-scoring domain. These values reflect the likelihood of observing the given alignment by chance; lower E-values indicate more significant matches. When both values are low, the sequence is likely homologous to the model. In some cases, however, only the full sequence E-value is significant, while the best domain E-value remains high. This situation may arise when a protein contains multiple domains, not all of which are Kunitz-like: the cumulative contribution of several weakly matching regions can lower the full sequence E-value, even if no individual domain strongly supports homology on its own. Such cases require careful interpretation, as they may also reflect remote homology with limited local similarity (Eddy, 2010 [2]). For this reason, evaluating both the full sequence and best domain E-values was essential. While the full sequence score captures the total matching signal across the entire protein, the domain-level score focuses on the most conserved region. This dual evaluation provided a more reliable assessment, particularly for ambiguous cases, multidomain proteins, and distant homologs, and allowed us to better balance sensitivity and specificity when identifying Kunitz domains.

For each sequence, the following information was extracted: UniProt ID, full sequence E-value and best domain E-value; it was also added a true class label (1 for Kunitz, 0 for non-Kunitz). This information was stored in .class files, one per subset. Some negative sequences were not included in the output due to extremely high E-values. To avoid biasing the evaluation, these entries were manually added back into the .class files with placeholder E-values of 10.0 and a class label of 0.

Positive and negative examples were then merged into two evaluation sets: set_1_strali.class and set_2_strali.class (set_1_seqali.class and set_2_seqali.class, which contained the same proteins as “strali” sets, but with e-value obtained using HMM built on MSA). These sets were used in a 2-fold cross-validation framework. For each model, the best classification threshold (E-value) was selected by testing values from 1e-01 to 1e-09 and choosing the one that maximized the Matthews Correlation Coefficient (MCC). Each threshold was then applied to the opposite set to evaluate model generalization, and a third evaluation was performed on the combined set to assess overall performance.

The script performance.py was used to compute standard classification metrics: accuracy (Q2), Matthews Correlation Coefficient (MCC), true positive rate (TPR) and precision (PPV). In addition, it was done a cross check to find false positives and false negatives. All these results were saved in hmm_results_strali.txt and hmm_results_seqali.txt for the structure-based and sequence-based models, respectively.

3. Results and discussion

3.1 MStA and MSA visualization on Jalview

The alignments obtained using MUSCLE (sequence-based) and PDBeFold (structure-based) were visualized with Jalview (Waterhouse et al., 2009 [4]) to inspect conserved residues, consensus regions, and sequence variability across the representative Kunitz domains (Figure 3-Figure 4). In both alignments, several highly conserved regions were observed, including the cysteine residues involved in the formation of disulfide bridges, a hallmark of the Kunitz fold.

A notable difference between the two alignments was found in the occupancy, which was generally higher in the sequence-based alignment. This is likely due to the fact that sequence alignments tend to introduce fewer gaps compared to structure-based alignments, which prioritize three-dimensional spatial correspondence over linear residue continuity (Carpentier et al., 2019 [15]).

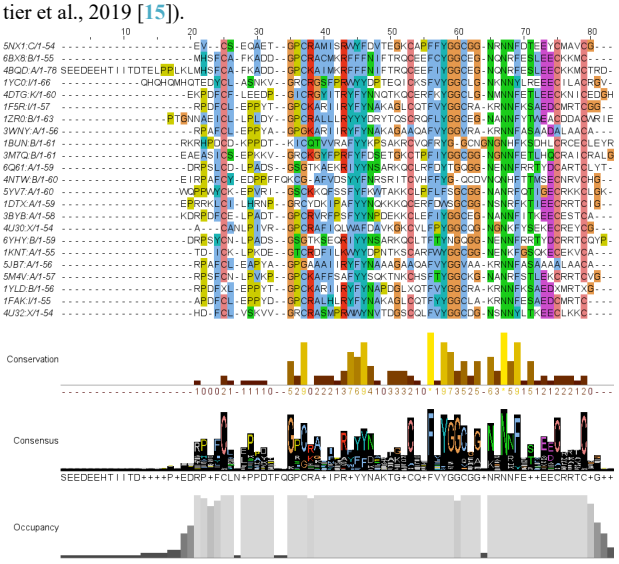


Figure 3 Multiple-structure alignment. The includes several annotation tracks: conservation (physicochemical similarity), consensus (most frequent residues, visualized with a sequence logo), and occupancy (number of sequences aligned at each position)(Martin et al., 2019 [3]). Conserved cysteines involved in disulfide bridges are clearly visible.

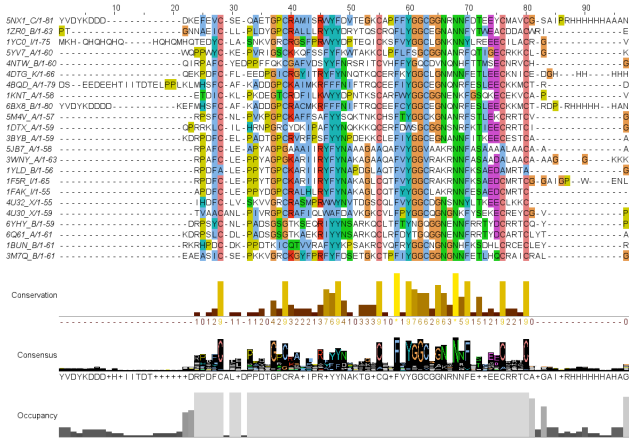


Figure 4 Multiple-sequence alignment. The alignment was constructed using the same representative sequences selected for the structure-based analysis. Annotation tracks include conservation, consensus, and occupancy, the latter reflecting the number of sequences with a non-gap residue at each position—generally higher here than in the structural alignment, due to fewer gaps introduced (Carpentier et al., 2019 [15]). Conserved cysteine residues, essential for forming the Kunitz domain's three disulfide bridges, are clearly visible and consistently aligned across all sequences also in this case.

3.2 Best threshold selection

An evaluation was performed across a range of E-value thresholds, from 1e-01 to 1e-09, to identify the values that yielded the best classification performance for both SET 1 and SET 2, considering both E-value sources: full sequence and best single domain. The relationship between the threshold and the Matthews Correlation Coefficient (MCC) was visualized in a series of plots (Figure 5).

The MCC is a balanced metric that considers true and false positives and negatives and is especially useful in evaluating binary classifiers on imbalanced datasets (as in this case where negatives are much more than positives). It is defined as:

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

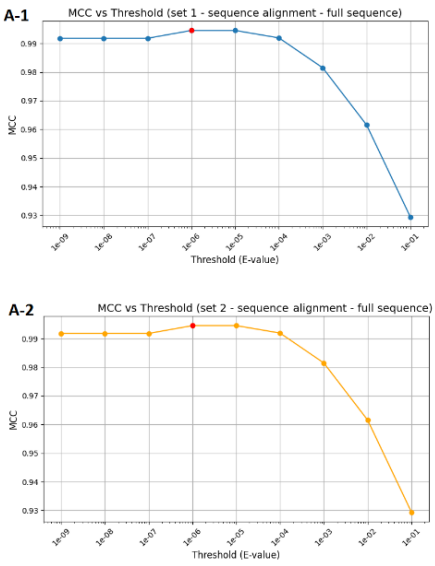
where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

This coefficient returns a value between -1 and 1, where 1 indicates perfect prediction, 0 no better than random, and -1 total disagreement between prediction and observation.

For each combination of set and E-value source, the threshold corresponding to the highest MCC was selected as the best threshold. In cases where multiple thresholds yielded the same MCC value, the lowest threshold (i.e., the most stringent) was selected. This conservative criterion favors minimizing false positives and ensures that the classifier remains selective, especially when performance is equivalent across multiple thresholds. This analysis was conducted for both the sequence-based and structure-based models.

For both methods the best threshold was the same (1e-05 for single-domain and 1e-06 for full-sequence) (Table 1-Table 2). This difference likely reflects the distinct scoring strategies: full-sequence scores capture cumulative alignment signals across the entire protein, including weakly conserved regions, and therefore require a more stringent threshold to avoid false positives. In contrast, single-domain scores focus on the most conserved local region, which may demand a slightly more relaxed threshold to maintain sensitivity—especially in the presence of multidomain proteins or remote homologs.



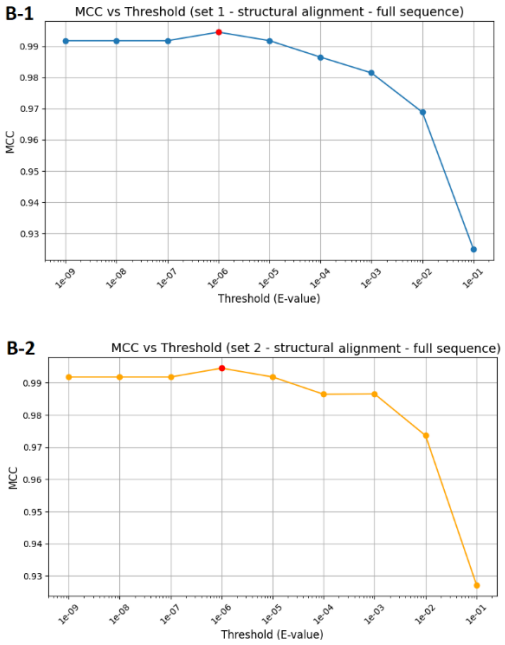


Figure 5 Graphical representations of the E-value thresholds ranging from 1e-01 to 1e-09. The red dot marks the best threshold, which was later used for cross-validation testing.

In (A), the plots show the MCC curves for SET 1 (blue) and SET 2 (yellow), using full-sequence E-values derived from the HMM built on the multiple sequence alignment (MSA).

In (B), the corresponding results are shown for the HMM constructed from the multiple structural alignment (MStA).

For clarity, only the full sequence MCC plots are shown in the main manuscript, while the domain-level results and plots can be accessed in the "graphs_and_tables/Best thresholds plots" folder of the GitHub repository.

Test set	E-value source	Threshold
set1	full-seq	1e-06
set2	full-seq	1e-06
set1	single-dom	1e-05
set2	single-dom	1e-05

Table 1 Best-performing E-value thresholds obtained for each test set and E-value scoring method, based on the HMM model built from the sequence-based alignment. There are two different appliable thresholds if the evaluation is on full-seq or single-domain e-value source.

Test set	E-value source	Threshold
set1	full-seq	1e-06
set2	full-seq	1e-06
set1	single-dom	1e-05
set2	single-dom	1e-05

Table 2 Best-performing E-value thresholds obtained for each test set and E-value scoring method, based on the HMM model built from the structure-based alignment. Like in sequence-based method there are two different appliable thresholds if the evaluation is on full-seq or single-domain e-value source.

3.3 Cross-validation results and case analysis

The results of the cross-validation phase are available in the folder OUTPUT EXAMPLE, under the corresponding hmm_results_strali.txt and hmm_results_seqali.txt files. For the structure-based HMM, the confusion matrices showed excellent classification performance across both cross-validation folds and the overall evaluation, with only a limited number of false positives and false negatives detected (Table 3).

Among the false negatives, the same four proteins were consistently observed: *D3GGZ8*, *O62247*, *A0A1Q1NL17*, and *Q8WPG5* (Table 4). These are all known Kunitz proteins, yet they exhibit considerable sequence and structural divergence. At the selected thresholds, the model failed to classify them correctly under both the full-sequence and single-domain scoring modes, likely due to their deviation from the canonical Kunitz profile.

The false positive results revealed a particularly interesting case: *P84555* (Table 4). This protein was classified as positive only under the single-domain scoring mode and appeared in all evaluations despite being labeled as negative in the dataset. Manual inspection confirmed that *P84555* (Table 4) is indeed a Kunitz-type protein, but it was excluded from the positive set because it lacked the PF00014 Pfam annotation and was therefore not retrieved by the UniProt query.

The sequence-based model identified the same set of false negatives and *P84555* as a false positive (Table 4). However, under the single-domain scoring, it failed to detect *Q11101* (Table 4), another known Kunitz protein correctly classified by the structure-based model. Despite having a Kunitz-like structure, *Q11101*'s sequence is likely too divergent for the sequence-aligned HMM to recognize it reliably, highlighting a limitation in the model's generalization ability when relying solely on sequence information.

Although the analysis could be considered complete at this point, further considerations were necessary. As it became apparent, the negative set included proteins that were, in fact, Kunitz, but not annotated with PF00014, and therefore were not removed during dataset preparation. To address this limitation, the analysis was repeated using an expanded annotation criterion: proteins annotated with the InterPro ID IPR036880, which also describes the Kunitz domain, were now included (results and material for this second analysis are available in the folder INTERPRO_ANALYSIS on the github).

The impact of this change was significant. Although overall MCC values decreased slightly, a deeper inspection of the false positives revealed that the models were now able to identify many more true Kunitz proteins that had neither PF00014 nor InterPro annotations. In particular, the structure-based model, using single-domain scoring, was able to correctly classify eight previously unannotated Kunitz proteins (*P84556*, *P85040*, *P85039*, *P0DJ63*, *P0DM47*, *P0DV02*, *P83604*, *P0DV07*) (Table 3-Table 4). The sequence-based model, while still solid, showed slightly lower sensitivity under the same conditions, correctly identifying only four of these eight proteins (*P84556*, *P85040*, *P85039*, *P0DJ63*) (Table 3-Table 4).

Interestingly, when using full-sequence scoring, both models performed similarly in terms of false positive and false negative detection. Nonetheless, under single-domain evaluation, the structure-based HMM demonstrated superior performance, especially in identifying Kunitz proteins affected by cross-reference annotation issues.

Model	E-value Type	TP	TN	FP	FN	Avg MCC
Only Pfam-based Evaluation						
Structure	Single-domain	364	572571	1	4	0.993182680
Structure	Full-sequence	364	572572	0	4	0.994546894
Sequence	Single-domain	363	572571	1	5	0.993180816
Sequence	Full-sequence	364	572572	0	4	0.994546894
Including InterPro-based Evaluation						
Structure	Single-domain	376	572552	8	4	0.984334276
Structure	Full-sequence	376	572556	4	4	0.989064052
Sequence	Single-domain	376	572556	4	4	0.989064052
Sequence	Full-sequence	376	572556	4	4	0.989064052

Table 3 Confusion matrices and average Mattew Correlation Coefficient for each model and E-value source type, divided by annotation method.

Model	E-value Type	False Negatives	False Positives
Only Pfam-based Evaluation			
Structure	Single-domain	D3GGZ8, O62247, A0A1Q1NL17, Q8WPG5	P84555
Structure	Full-sequence	D3GGZ8, O62247, A0A1Q1NL17, Q8WPG5	NA
Sequence	Single-domain	D3GGZ8, O62247, A0A1Q1NL17, Q8WPG5, Q11101	P84555
Sequence	Full-sequence	D3GGZ8, O62247, A0A1Q1NL17, Q8WPG5	NA
Including InterPro-based Evaluation			
Structure	Single-domain	Q8MVZ2, Q8MVZ3, P36235, D3GGZ8	P84556, P85040, P85039, P0DJ63, P0DM47, P0DV02, P83604, P0DV07
Structure	Full-sequence	Q8MVZ2, Q8MVZ3, P36235, D3GGZ8	P84556, P85040, P85039, P0DJ63
Sequence	Single-domain	Q8MVZ2, Q8MVZ3, P36235, D3GGZ8	P84556, P85040, P85039, P0DJ63
Sequence	Full-sequence	Q8MVZ2, Q8MVZ3, P36235, D3GGZ8	P84556, P85040, P85039, P0DJ63

Table 4 False positives and false negatives identified across models, E-value source types, and annotation methods (Pfam and InterPro).

4. Conclusion

Overall, the structure-based HMM performed better than the sequence-based model, as expected. Nevertheless, the difference in performance was not substantial— the sequence-based HMM also proved to be highly robust and reliable across multiple evaluations.

Importantly, both models played a key role in uncovering cross-reference errors in domain annotation, successfully identifying Kunitz proteins that had been incorrectly excluded from the positive set due to missing Pfam or InterPro labels. This highlights a critical limitation of relying solely on annotation-based filtering and underscores the value of model-based inference.

A key aspect worth highlighting is the MCC. This metric is widely regarded as a reliable indicator for evaluating the performance of HMM-based classifiers. As shown in Table 3–Table 4, the MCC suggests that the structure-based model outperforms the sequence-based one when using the single-domain E-value. Indeed, the structure-based HMM correctly classifies Q11101, a known Kunitz protein with significant sequence divergence, which the sequence-based model fails to detect—highlighting its lower flexibility in handling sequence variation.

However, when evaluating the InterPro-augmented dataset, overall MCC values drop significantly. In this case, it is crucial to avoid hasty conclusions and instead inspect the nature of the reported false positives. Upon manual inspection, many of these proteins were found to be true Kunitz proteins that had been incorrectly retained in the negative set due to missing both the PF00014 and IPR036880 annotations.

Contrary to initial expectations, the inclusion of InterPro annotations led to better detection of proteins affected by cross-reference errors—especially by the structure-based model, which identified more unannotated Kunitz proteins than its sequence-based counterpart. While this increased the number of reported "false positives", it actually reflects greater generalization capacity, not poorer performance.

Thus, although MCC remains a useful and informative metric, relying on it without considering dataset limitations, annotation inconsistencies, or systematic biases can lead to misleading interpretations. Proper evaluation must always be supported by critical inspection of classification outcomes. This observation not only justifies the lower MCC in that case but also reinforces the higher generalization ability of the structure-based model,

especially in detecting domain instances beyond the boundaries of existing annotations.

Therefore, this work also serves as a reminder of the importance of conscious validation and critical evaluation of automated results. Apparent misclassifications may, in fact, point to errors in reference datasets or reveal biologically relevant signals missed by annotation pipelines.

In conclusion, while both models demonstrated strong performance and the ability to accurately identify Kunitz proteins, they remain open to further improvement—particularly in enhancing flexibility and reducing false negatives. Future work should aim to continuously refine the positive set by identifying and recovering Kunitz proteins that are currently misclassified due to annotation limitations.

Acknowledgements

This work was carried out with the support of Professor Emidio Capriotti, Associate Professor at the University of Bologna, during the course "Laboratory of Bioinformatics 1" under his supervision.

Special thanks also go to the open-source tools used throughout this project. In particular, I acknowledge the use of HMMER, MUSCLE, PDBeFold, and Jalview, which were essential for the successful development and execution of the analyses.

References

1. Ascenzi P, Bocedi A, Bolognesi M, et al. The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein. *Current Protein & Peptide Science*. 2005;4(3). doi:10.2174/1389203033487180
2. Eddy SR. *HMMER User's Guide Biological Sequence Analysis Using Profile Hidden Markov Models*.; 2010. http://hmmer.org/Version3.0rc2;http://eddylab.org/
3. Martin D, Procter J, Waterhouse A, et al. *Jalview 2.11 Manual and Introductory Tutorial*.; 2019.
4. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9). doi:10.1093/bioinformatics/btp033
5. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23). doi:10.1093/bioinformatics/bts565
6. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004;32(5). doi:10.1093/nar/gkh340
7. Edgar RC, Batzoglou S. Multiple sequence alignment. *Current Opinion in Structural Biology*. 2006;16(3). doi:10.1016/j.sbi.2006.04.004
8. Bystroff C, Krogh A. Hidden Markov Models for Prediction of Protein Features. In: *Protein Structure Prediction*. ; 2008. doi:10.1007/978-1-59745-574-9_7
9. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*. 2009;10(6). doi:10.2174/138920209789177575
10. Laskowski M, Kato I. Protein inhibitors of proteinases. *Annual review of biochemistry*. 1980;49. doi:10.1146/annurev.bi.49.070180.003113
11. Creighton TE. *The Two-Disulphide Intermediates and the Olding Pathwa of Reduced Pancreatic Trypsin Inhibitor*. Vol 95.; 1975.
12. Kunitz M, Northrop JH. Isolation from beef pancreas of crystalline trypsinogen, trypsin, a trypsin inhibitor, and an inhibitor-trypsin compound. *Journal of General Physiology*. 1936;19(6). doi:10.1085/jgp.19.6.991
13. Chand HS, Schmidt AE, Bajaj SP, Kisiel W. Structure-Function Analysis of the Reactive Site in the First Kunitz-type Domain of Human Tissue Factor Pathway Inhibitor-2. *Journal of Biological Chemistry*. 2004;279(17). doi:10.1074/jbc.M400802200
14. Chua LM, Lim ML, Wong BS. The Kunitz-protease inhibitor domain in amyloid precursor protein reduces cellular mitochondrial enzymes expression and function. *Biochemical and Biophysical Research Communications*. 2013;437(4). doi:10.1016/j.bbrc.2013.07.022
15. Carpentier M, Chomilier J, Valencia A. Protein multiple alignments: Sequence-based versus structure-based programs. *Bioinformatics*. 2019;35(20):3970-3980. doi:10.1093/bioinformatics/btz236
16. Fratini E, Rossi MN, Spagoni L, et al. Molecular Characterization of Kunitz-Type Protease Inhibitors from Blister Beetles (Coleoptera, Meloidae). *Biomolecules*. 2022;12(7). doi:10.3390/biom12070988