Bioinformatics, 2025, 1-6
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 23 November 2025
Manuscript Category

Report

Profile Hidden Markov Model for the BPTI/Kunitz domain annotation

Domenico Zianni[1,*]

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum–Università di Bologna, Via F. Selmi 3-4th Floor, Room 183

*To whom correspondence should be addressed.

Associate Editor: Emidio Capriotti

Abstract

Motivation: The BPTI/Kunitz domain is a cysteine-rich protease-inhibitor motif whose strong structural conservation makes it ideal for profile-based detection. This project aims to develop and compare sequence and structure-derived HMMs to identify the domain reliably in protein datasets. The goal is to improve sensitivity and specificity in protein annotation by exploiting both evolutionary and structural signals

Results: The results highlight both the robustness of HMM-based detection, which achieved consistently high performances (MCC > 0.99) across different E-value threshold. The study showed the importance of training alignments to minimize misclassification while preserving sensitivity.

Availability: All data, scripts and graphs are available at GitHub repository of the author

Contact: domenico.zianni@studio.unibo.it

1. Introduction

Kunitz domain is a type of active domain found in proteins that behaves as protease inhibitor. A commonly used model for this family is aprotinin (bovine pancreatic trypsin inhibitor) but the family includes numerous other members such as snake venom basic protease (mammalian inter-alpha-trypsin inhibitors), trypstatin (a rodent mast cell inhibitor of trypsin) and tissue factor pathway inhibitor precursor (TFPI). Bovine pancrease transferase inhibitor (BPTI) is the first Kunitz-active domain isolated and crystallized that inhibits the function of proteases (3), It is about 50–60 amino acids long with a molecular weight of about 6 kDa folded into a disulfide-rich α/β structure constrained by 3 disulphide bonds (Fig.1). It is characterized by conserved spacing between cysteine residues. BPTI Kunitz acts as serin-protease inhibitor (1,2). Kunitz inhibitors directly block the serine protease active site without any conformational change, forming an anti-parallel β sheet between enzyme and inhibitor. The binding kinetics and enzyme affinities vary with each invertebrate Kunitz inhibitor. The segment responsible for protease inhibition is called the protease-binding loop.

This convex, extended and solvent-exposed loop is highly complementary to the concave active site of the enzyme. In the complex of protease–inhibitor interaction, 10–17 amino acid residues on the inhibitor site and 17–29 residues of the protease make numerous van der Waals and hydrogen bond interactions. It is generally recognized that the N-terminal side of the reactive site (P) is energetically more important than the C-terminal side. Stand-alone Kunitz domains have been used as a framework for the development of new pharmaceutical drugs. BPTI is an extensively studied model protein belonging to this family.

Hidden Markov Models (HMMs) are statistical frameworks that treat a macromolecular sequence as a series of observable symbols generated by a chain of unobserved (“hidden”) states obeying the Markov property. By explicitly modelling position-specific dependencies (for example, the tendency of certain residues or motifs to co-occur) HMMs excel at capturing the subtle signals embedded in proteins, such as conserved

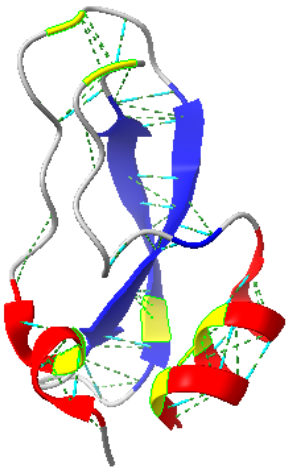


Figure 1; Crystal Structure of BPTI (PDB: 1BPTI)

UCSF ChimeraX (version 1.9) was used to render the structure, α helices (red) and β strands (blue) are distinguished, cysteine residues are labelled with yellow, Hydrogen bonds (blue) and weaker molecular interactions (Van der Waals in green) are represented by dashes.

active-site loops or structural cores (4). A specialized subclass, the profile HMM, adopts a linear left-to-right architecture in which each position in a multiple sequence alignment is represented by a triplet of states (match, insert, delete). This design enables the model to learn both fold-specific patterns (common to all family members) and function-specific motifs (unique to subfamilies), yielding highly sensitive detectors of remote homology and motif occurrence (4,10).

In this study, a profile HMM was built for the BPTI/Kunitz protease-inhibitor domain by first structurally aligning a diverse set of experimentally determined Kunitz domains. The resulting model parameters (transition probabilities encoding the typical length and gap patterns of the domain and emission probabilities reflecting the conserved cysteines and active-site residues) allowed to scan large sequence databases and reliably identify and annotate Kunitz domains even when overall sequence identity is low.

Furthermore, by incorporating a curated set of negative sequences into training, the output thresholds were fine-tuned to minimize false positives arising from general fold signals, thereby maximizing the functional specificity of our detector. Ultimately, this profile HMM provides a sensitive and specific tool for both protein family classification and domain annotation in proteomes of interest.

2. Material and Methods

The complete workflow of the project including the coding hands-on and the packages used to process the data (CD-HIT, BLAST+, HMMER, Biopython) is available at the associated GitHub repository cited above.

2.1 Data gathering

A representative set of protein structures used to perform the HMM was retrieved from RCSB PDB (6) advanced search selecting the following constraints:

- Data collection resolution ≤ 3.5Å
- Polymer Entity Sequence Length ≥ 45
- Polymer Entity Sequence Length ≤ 80
- Identifier - Pfam Protein Family is PF00014

PFAM identifier serves as describer of Kunitz BPTI family features.
The resulting hits of the research (160) were collected in a custom report csv file that was converted to fasta in order to perform the clustering.

2.1.1 Clustering

The process was performed using CD-HIT (5) (version 4.8.1), a fast program for clustering and comparing large sets of protein or nucleotide sequences with 90% of identity threshold that allowed the identification of sets of proteins (25 clusters) on which to perform the multiple structural alignment while avoiding redundancy. The text file returned from the clustering was used to select a representative structure for each cluster according to the sequence length and identity. The representative dataset was cleaned to ensure that no bias caused by unusual sequence length or misleading character was introduced in the alignment.
For cleaning process, the following criteria were set:

- Sequence length: between 40 and 100 residues (taken as safe range since Kunitz domain = 50/60 residues)
- Unknown characters warning

Two entries which did not respect length constraints were removed: 2ODY:E (127 residues) and 5JBT:Y (38 residues).
Another entry was manually removed (4BQD:A) because its A chain lacks one or both β-strands typical of Kunitz fold; likely a partial or degraded domain rather than a full, functional one.
The cleaned sequences were converted into a FASTA file with a bash pipeline

2.2 Multiple Structural Alignment

The structural alignment was calculated using PDBeFold (7), a well-known tool in Bioinformatics to evaluate structure similarity. The following filters have been selected:
• Submission form/3D alignment: multiple
• Source: List of PDB codes
The list of representative PDB IDs was used as a query. The returned alignment was double-checked to highlight possible errors; Root Mean Square Deviation (RMSD) was evaluated to measure the difference between structures, Secondary Structure Elements (SSEs) since they're crucial in protein fold and Q-score,

which takes into account the number of residues in the matched SSEs and their positions in space. After launching the PDBeFold, the alignment was saved and exported in a FASTA file.

2.3 HMM: Building and Testing

Profile HMM was generated starting from the structural alignment with *hmmbuild* command from HMMER software package that provides tools for making probabilistic models of protein and DNA sequence domain families (profile HMMs), and for using these profiles to annotate new sequences, to search sequence databases for additional homologs, and to make deep multiple sequence alignments (8).

The positive testing set was created by performing an advanced search on UniProt using the Pfam ID for the BPTI/Kunitz domain (00014) and filtering the results to entries in Swiss-Prot, which are manually curated and of higher quality. This will yield a total of 380 proteins containing the BPTI/Kunitz domain, which will serve as positive set. All remaining Swiss-Prot proteins that do not contain the Kunitz domain were used as negative set.
To ensure a fair evaluation of HMM, it's important to remove from the positive set any sequences that are too similar to the structures used to build your model in this way it's possible to prevent any overlap between training and testing data.

Using BLAST's (version 2.12.0+) (10) makeblastdb a BLAST database was created with the Kunitz proteins. With blastp, the 22 representative sequences used for the multiple structural alignment were searched against the Kunitz database.

At this point balanced positive and negative test sets for Kunitz protein validation were prepared by removing redundant sequences, shuffling, splitting, and extracting the relevant FASTA entries.
All sequences within defined thresholds (sequence identity ≥ 95% and number of aligned residues ≥ 50) were removed from the positive dataset to avoid biases in the validation, for a final amount of 365 positive sequences.

Positive set was shuffled and splitted in two different sets, pos_set_1(183 IDs) and pos_set_2 (182 IDs). pos_set_1 contained one more entry since the whole Kunitz dataset had odd number of entries, this small difference does not affect the validity of the evaluation, as long as the split is random and unbiased. Entries in both positive sets were extracted from the whole Uniprot/SwissProt database (release 2025_01)

Negative testing set was built extracting all Uniprot IDs from SwissProt dataset and removing all Kunitz IDs to get negative candidates. Then the retrieved IDs were shuffled and splitted in two subsets: neg_set_1 (286,286 IDs) and neg_set_2 (286,286 IDs).
Sequences from both sets were extracted from SwissProt and stored in FASTA format.
For both positive and negative testing set SwissProt database was used as a background model, providing non-Kunitz protein sequences for negative test sets in order to assess the specificity of the constructed profile HMM.

The hmmsearch program was used to evaluate the trained profile by searching it against both the positive (Kunitz) and negative (non-Kunitz) sequence databases. This allowed assessment of the model's sensitivity and specificity (11). After that, HMM results for each sequence were stored in classification files contained labelled positive and negative entry (0 for negative, 1 for positive) and relevant full-sequence and best domain E-values.
To prevent negative sequences underrepresentation, negative IDs in neg_set_1 and neg_set_2 that are missing from the corresponding classification files were found and added to the .class files with a fake high E-value (10.0), indicating no detection.
Final dataset was created with combination of results into final classification files for further analysis.

3. Results and discussion

3.1 Multiple Sequence Alignment and Structural Alignments

To explore sequence and structural conservation within the Kunitz-type protease inhibitor dataset, both multiple sequence and structural alignments were visualized using dedicated bioinformatics tools.
Multiple sequence alignments (MSA) obtained from PDBeFold, and HMM-based analyses were examined using Jalview (12) (Fig.2), which provides a graphical representation of sequence conservation, residue variability, and secondary structure prediction. Strong conservation peaks correspond to the six cysteine residues forming the disulfide-bond framework and to residues in the β-sheet core, reflecting the high structural constraints required to maintain the Kunitz fold (13). Alignment quality is highest in these central regions, confirming that match positions are consistently and reliably aligned across all sequences. In contrast, lower conservation and lower quality scores occur in the N- and C-terminal extensions and in flexible loop regions, where insertions and substitutions are more frequent. The consensus sequence and sequence logo reinforce this pattern, showing dominant cysteine and aromatic residues at core positions and greater variability in surface-exposed regions. Occupancy remains high across the structured core but drops at the endings, reflecting differences in signal peptides and peripheral segments. Overall, the MSA reveals a well-defined, highly conserved domain core suitable for profile HMM

construction, with peripheral variability that does not disrupt the essential Kunitz architecture.

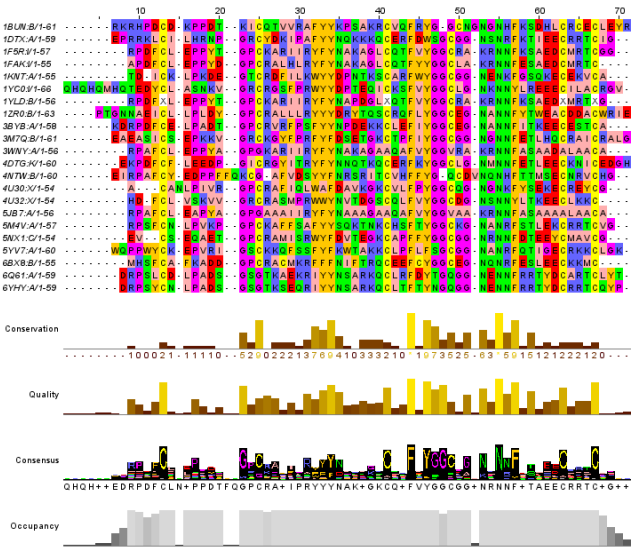


Figure 2: *The alignments were coloured according to residue conservation and biochemical properties (Zappo colouring scheme), and consensus sequences were computed to identify conserved motifs among homologs.*

Structural superposition generated by PDBeFold were visualized in UCSF ChimeraX (9) (Fig.3), allowing a three-dimensional inspection of conserved structural cores and local deviations among aligned protein models.

Combined use of Jalview and ChimeraX enabled an integrated analysis of sequence–structure relationships, providing insight into conserved residues potentially involved in functional or stabilizing roles within the Kunitz protein family.

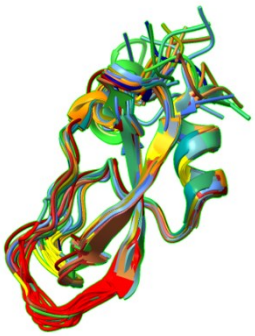


Figure 3: *The matched residues were coloured according to conservation scores derived from the sequence alignment, and pairwise RMSD values were evaluated to assess structural similarity.*

3.2 HMM Performance Evaluation

To evaluate the ability of the HMM to correctly identify proteins containing the BPTI/Kunitz domain, a confusion matrix was built.

The confusion matrix is essential because it separates correct and incorrect predictions into true positives, true negatives, false positives, and false negatives. This decomposition allows a more detailed assessment than accuracy alone and forms the basis for all subsequent performance metrics.

From the confusion matrix, the following measures were calculated:

- Q2 (overall accuracy): proportion of correctly classified sequences.
- TPR (True Positive Rate / Sensitivity): fraction of real Kunitz-domain proteins correctly detected.
- PPV (Positive Predictive Value / Precision): fraction of predicted positives that are truly positive.

The main metric used to judge model quality is the Matthews Correlation Coefficient (MCC), defined as:

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

Where:

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

MCC was chosen because it remains informative even when the positive and negative datasets are highly unbalanced, as is the case here (few Kunitz-domain proteins vs. the large SwissProt background). Unlike accuracy or TPR, MCC incorporates all four entries of the confusion matrix and penalizes false positives and false negatives symmetrically, making it a robust indicator of the model's discriminative power (13).

To assess how the HMM discriminates Kunitz-domain proteins from non-Kunitz sequences, classification performance was evaluated using two scoring strategies: the full-sequence E-value and the best-domain E-value. These two measures capture different biological scenarios. The full-sequence score reflects how confidently the model matches the entire protein, which is useful when domain boundaries are clear and the domain occupies most of the sequence. However, many proteins contain multiple domains or long unstructured regions, and in those cases the global score may be diluted. The best-domain E-value compensates for this by evaluating only the most convincingly aligned domain-sized region, making it more sensitive to detecting single Kunitz domains embedded in large or multidomain proteins. Using both approaches increases coverage and robustness but requires careful threshold selection and interpretation to avoid false positives/negatives.

For both strategies, performance was measured across decreasing E-value thresholds (10^{-1} to 10^{-12}). As shown in Figure 5, the MCC increases sharply at

stricter thresholds, reaching a plateau between 10^{-5} and 10^{-8}, where both test sets achieve values close to 1.0. This indicates a near-perfect balance between true and false predictions in this range.

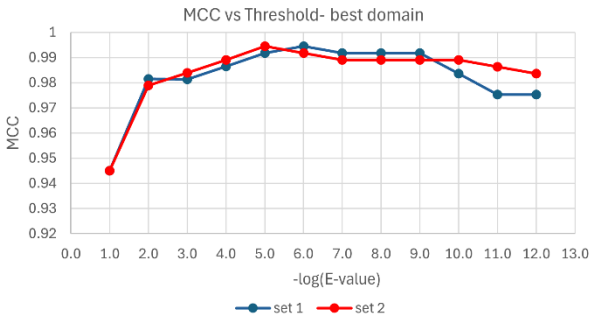
False positives (FP) and false negatives (FN) arise from different biological and algorithmic sources:

False positives mainly occur at looser E-value thresholds (10^{-1} to 10^{-3}). Proteins without a Kunitz domain may still show weak similarity in short, compositionally biased regions, leading the model to classify them as positives. These cases highlight the trade-off between sensitivity and specificity: relaxing the threshold increases the detection of true Kunitz proteins but also "admits" unrelated sequences with accidental similarity.

False negatives appear when thresholds become too strict (10^{-10} to 10^{-12}). Here, native Kunitz proteins with sequence divergence fall below the score cutoff and are incorrectly rejected. These are biologically real variants that remain structurally compatible with the Kunitz fold but are penalized by the HMM due to alignment uncertainty or low emission probabilities at specific positions.

Comparing the two evaluation modes, the best-domain E-value generally reduces false negatives, because the classification focuses on the domain region alone rather than the whole protein sequence, which may contain long disordered or unrelated regions that depress the full-length score. In contrast, full-sequence E-value is more susceptible to both FP and FN, especially for long multidomain proteins where the Kunitz region represents only a small fraction of the total length.

Globally, the trends observed suggest that classification performance depends critically on selecting an appropriate E-value threshold and on the quality of the training alignment, refining the MSA could further reduce both FP and FN rates and increase robustness across thresholds.



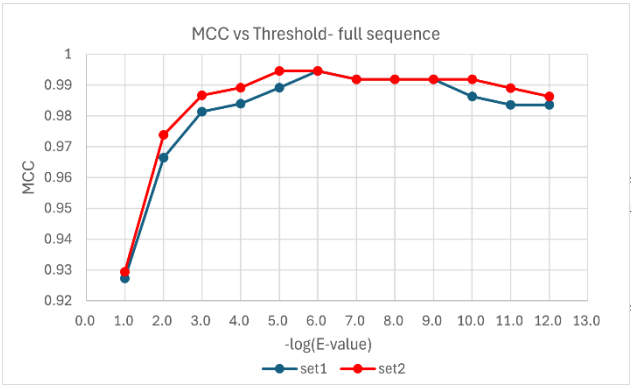


Figure 4 *Variation of the MCC for best-domain and full-sequence classification performance across decreasing E-value thresholds (from 1e-1 to 1e-12). The evaluation shows set 1 (blue) and set 2 (red) MCC behaviour.*

Table 1. Best-performing threshold for both testing set of full sequence and best domain evaluation

Test set	Type	Best-Threshold
1	Full sequence	1e-6
2	Full sequence	1e-6
1	Best domain	1e-6
2	Best domain	1e-5

Conclusions

This study demonstrated that profile Hidden Markov Models are highly effective tools for detecting the BPTI/Kunitz domain, a structurally conserved inhibitor fold characterized by a stable β-sheet core and cysteine-rich framework. By integrating structural information into the training alignment and evaluating performance across different E-value thresholds, the models achieved consistently high classification accuracy, with MCC values approaching 0.99 for both full-sequence and best-domain searches. The comparison between the two evaluation strategies confirmed that domain-level scoring captures true positives that could otherwise be overlooked by large non-domain regions, while full-sequence scoring remains valuable for identifying globally conserved architectures. The few false positives observed were associated with cysteine-rich or disulfide-stabilized proteins that partially resemble Kunitz-like motifs, whereas false negatives arose mainly from sequences exhibiting atypical loop lengths or partial domain truncations.

Acknowledgements

This project was carried out under the guidance of prof. Emidio Capriotti during Laboratory of Bioinformatics 1 course at the University of Bologna

References

1. Shiwanthi Ranasinghe and Donald P. McManus, “STRUCTURE AND FUNCTION OF INVERTEBRATE KUNITZ SERINE PROTEASE INHIBITORS,” *Developmental & Comparative Immunology* 39, no. 3 (March 1, 2013): 219–27, https://doi.org/10.1016/j.dci.2012.10.005;

2. Amy E. Schmidt et al., “Crystal Structure of Kunitz Domain 1 (KD1) of Tissue Factor Pathway Inhibitor-2 in Complex with Trypsin: IMPLICATIONS FOR KD1 SPECIFICITY OF INHIBITION*,” *Journal of Biological Chemistry* 280, no. 30 (July 29, 2005): 27832–38, https://doi.org/10.1074/jbc.M504105200.

3. M. Kunitz and John H. Northrop, “ISOLATION FROM BEEF PANCREAS OF CRYSTALLINE TRYPSINOGEN, TRYPSIN, A TRYPSIN INHIBITOR, AND AN INHIBITOR-TRYPSIN COMPOUND,” Journal of General Physiology 19, no. 6 (July 20, 1936): 991–1007, https://doi.org/10.1085/jgp.19.6.991.

4. S R Eddy, ‘Profile Hidden Markov Models.’, Bioinformatics, 14.9 (1998), pp. 755–63, doi:10.1093/bioinformatics/14.9.755.

5. Alperen Degirmenci, Introduction to Hidden Markov Models, n.d

6. RCSB Protein Data Bank, “RCSB PDB: Homepage,” accessed June 19, 2025, https://www.rcsb.org/.

5."CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences", Weizhong Li & Adam Godzik. Bioinformatics, (2006) 22:1658-1659

7. “PDBe < Fold < EMBL-EBI,”, https://www.ebi.ac.uk/msd-srv/ssm/.

8. “HMMER,”biosequence analysis using profile hidden Markov models, http://hmmer.org/.

9. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. Protein Sci. 2021 Jan;30(1):70-8.

10. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215, no. 3 (1990): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

11. Eddy, Sean R. HMMER User’s Guide. n.d.

12. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189-91. doi: 10.1093/bioinformatics/btp033. Epub 2009 Jan 16. PMID: 19151095; PMCID: PMC2672624.

13. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE 12(6): e0177678. https://doi.org/10.1371/journal.pone.0177678

14. Bode, W. and Huber, R. (1992) Natural Protein Proteinase Inhibitors and Their Interaction with Proteinases. European Journal of Biochemistry, 204, 433-451. http://dx.doi.org/10.1111/j.1432-1033.1992.tb16654.