Bioinformatics, 2025, 06–22
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 22 June 2025

Bioinformatics

Structure-Based HMM Profiling for the Detection and Annotation of Kunitz-Type Protease Inhibitor Domains in SwissProt Sequences

Mahan Balooei
Bioinformatics Master's Degree Course, University of Bologna, Italy

Abstract

Motivation: The Kunitz-type protease inhibitor domain is a conserved structural motif critical for regulating protease activity across diverse species. Accurate detection is vital for understanding protein function. In this project, I developed two profile Hidden Markov Models (HMMs) for identifying Kunitz domains in protein sequences: one trained on a standard multiple sequence alignment (MSA) using MUSCLE, and another using a structure-based alignment derived from PDB data. Both models were trained on curated UniProt datasets and evaluated through cross-validation using Matthews Correlation Coefficient (MCC). While both models performed strongly (MCC ≥ 0.99), the structure-based HMM demonstrated superior sensitivity in detecting distant homologs and identifying true Kunitz domains not annotated in Pfam. These results were further validated against InterPro entries. The findings emphasize the added value of structural information in improving domain detection and annotation accuracy.
Availability: All code, data, and supplementary materials are available at Github Repository.

Contact: mahan.balooei@studio.unibo.it

1 Introduction

1.1 The Kunitz Domain

The Kunitz-type protease inhibitor (KPI) domain is a small protein motif that's about 50-70 amino acids long. What makes it special is its compact α+β fold that's held together by three disulfide bonds arranged in a C1-C6, C2-C4, and C3-C5 pattern. These disulfide bonds are really important - they keep the structure stable while allowing it to bind tightly to serine proteases without changing the protease's shape.

The domain was first found in bovine pancreatic trypsin inhibitor (BPTI), but we now know it's everywhere in nature. You can find Kunitz domains in mammals, parasites, and even venomous animals like snakes and scorpions. They're basically really good at stopping proteolytic activity, which is important for things like blood clotting, immune responses, and defense mechanisms.

From a medical perspective, these domains are pretty exciting. They could be used to develop new drugs - everything from protease inhibitors for treating inflammatory diseases and cancer to new anticoagulants for surgery. What's interesting is that some Kunitz domains are found in human proteins too. For example, there's a Kunitz-containing version of the amyloid precursor protein (APP) that might be involved in Alzheimer's disease.

Sometimes you'll find just one Kunitz domain in a protein, but other times there are multiple domains in tandem, which probably helps them target different proteases. Because they're so conserved and therapeutically relevant, we need good computational tools to find them. That's where profile Hidden Markov Models (HMMs) come in, and why databases like Pfam (PF00014) and InterPro (IPR002223) are so useful.

1.2 Hidden Markov Models

Hidden Markov Models (HMMs) are statistical tools that work really well for analyzing biological sequences like proteins and DNA. Profile HMMs are especially good at capturing both the conserved and variable parts of protein domains, which makes them essential for finding domains, annotating sequences, and classifying protein families.

A profile HMM learns from a multiple sequence or structure alignment of known domain examples. It can model which amino acids are likely to appear at each position, handle insertions and deletions, and account for evolutionary changes. The HMM has hidden states (representing things like amino acid positions in a domain) connected by transition probabilities, and emission probabilities that tell you how likely you are to see specific amino acids from each state.

This probabilistic approach is great for finding distant relatives of known proteins, even when the sequence similarity is low or annotations are missing. For this project, I used the HMMER suite to build a profile HMM for the BPTI/Kunitz domain. By including structural alignment data, I hoped to improve both sensitivity and specificity of Kunitz domain detection, especially given how these domains can vary across species.

HMMs are used in lots of bioinformatics tools like Pfam, SMART, and InterPro, where they're fundamental for automated protein domain annotation.

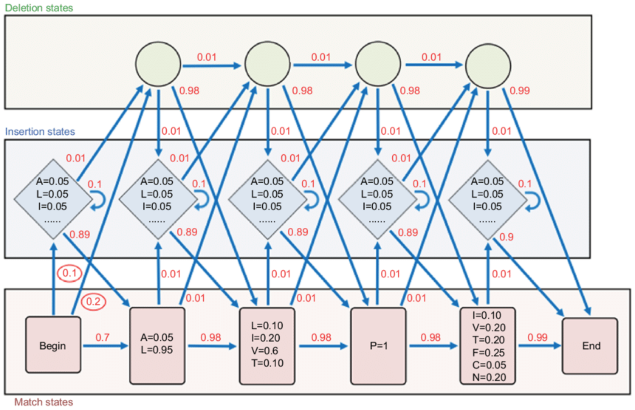


Diagram of a profile Hidden Markov Model (HMM): Match states (red rectangles), deletions (green circles), and insertions (blue diamonds). Red numbers show transition probabilities; equalities indicate emission probabilities without pseudocounts. Match states use alignment-based emissions; insertions use uniform background (1/20). Transition probabilities in red circles are discussed in the text; others are simplified for clarity.

2. Materials and Methods

2.1 Dataset Collection and Preprocessing

To build and test my HMMs for Kunitz domains, I collected both sequence and structure data from UniProtKB/Swiss-Prot and the Protein Data Bank (PDB).

For the sequence data, I downloaded protein sequences from UniProtKB/Swiss-Prot (accessed April 1st, 2025) using these filters: only human proteins (*Homo sapiens*, NCBI Taxonomy ID: 9606), reviewed entries only, and Pfam annotation PF00014. This gave me three datasets: human proteins with Kunitz domains (HK), human proteins without the domain (HNK), and non-human proteins with Kunitz domains (NHK). I made sure the validation sequences didn't overlap with my training data. I also downloaded the full Swiss-Prot database (March 18th, 2025) to use as background for creating negative control datasets.

For the structural modeling, I got Kunitz domain-containing protein structures from the PDB (accessed April 7th and 10th, 2025). I used an advanced search with these parameters: Pfam annotation PF00014, sequence lengths between 45-80 amino acids, and resolution ≤ 3.5 Å. This gave me 158 entries, which I processed to extract amino acid sequences.

To reduce redundancy and avoid bias in the alignments, I used CD-HIT (v4.8.1) with a 90% identity threshold. This gave me 25 clusters, and I picked a representative sequence from each cluster. I also removed sequences with weird lengths (like 2ODY:E) after checking them manually. The final non-redundant set was saved as pdb_kunitz_rp.fasta and used for multiple sequence and structural alignments.

All the scripts, intermediate files, and final datasets are documented and available in my GitHub repository.

2.2 Structure-Based Multiple Sequence Alignment

I performed structure-based multiple sequence alignment of the clustered Kunitz domain sequences using **PDBeFold** to check structural similarity and filter out any outliers. The alignment results came in .ali format, which I converted to FASTA and formatted for HMMER using custom AWK scripts.

To assess alignment quality, I used **RMSD** (Root Mean Square Deviation) with a filtering threshold of RMSD < 1.0. From the initial 24 aligned structures, I had to exclude one entry (PDB ID: 5JBT:Y) because it had a really high RMSD value of 2.918 Å, way above the dataset average (usually around 0.4-0.5 Å). I also looked at **Q-score** and **secondary structure elements (SSEs)** to make sure the alignment quality was good.

I used **UCSF Chimera (v1.20)** for visualization and manual inspection. The final alignment had 23 representative and structurally consistent sequences, saved as pdb_kunitz_rp_strali.fasta for building the profile HMM.

2.3 Preparation of Evaluation Sets and Cross-Validation

To test how well my trained HMMs performed, I prepared positive and negative datasets. The positive dataset combined human (HK) and non-human (NHK) Kunitz domain sequences from UniProtKB/Swiss-Prot, giving me 397 sequences (AK).

I created a **BLAST database** using makeblastdb (v2.16.0+) and searched the sequences I used in multiple sequence alignment against AK using **BLASTp**. I removed sequences with ≥95% identity and ≥50 aligned residues to prevent overlap with training data, ending up with 368 final positive sequences.

For the negative dataset, I used all non-Kunitz sequences (excluding any with Pfam ID PF00014) from the full UniProt/Swiss-Prot database - that's 572,573 sequences! I randomized both datasets and split them into two subsets each, creating four evaluation sets: **pos_1**, **pos_2**, **neg_1**, and **neg_2**. These were used in **2-fold cross-validation** to make sure the model would generalize well and avoid overfitting.

I evaluated using the hmmsearch tool, saving results in .out and .class formats with sequence ID, class label, bit score, and E-value. Each evaluation round involved merging one positive and one negative subset to form a complete test set.

2.4 HMM Evaluation and Performance Metrics

I built the profile HMMs from both sequence- and structure-based alignments using **HMMER's** hmmbuild **(v3.4)** and evaluated them using hmmsearch on the four validation subsets. For each sequence, I extracted the **bit score**, **full sequence E-value**, and **best domain E-value**, along with the true class label (1 for Kunitz, 0 for non-Kunitz), and saved everything in .class files. When negative sequences were missing from the output due to high E-values, I manually added placeholder values (E = 10.0) to avoid bias.

I assessed each model's performance by merging one positive and one negative set, trying different **E-value thresholds**(from 1e-01 to 1e-09), and picking the one that gave the highest **Matthews Correlation Coefficient (MCC)**. Then I tested this threshold on the opposite validation set to see how well it generalized, followed by evaluation on the combined dataset.

I wrote a custom Python script (performance.py) to calculate standard classification metrics: **Q2 (accuracy)**, **TPR (sensitivity)**, **PPV (precision)**, and **MCC**. I focused on MCC for threshold optimization because it's robust for imbalanced datasets and accounts for all true/false positive and negative rates. I also analyzed the results to identify **false positives and false negatives** and stored them in hmm_results_strali.txt and hmm_results_seqali.txt for the structure- and sequence-based models, respectively.

Looking at both full sequence and domain-level E-values helped with interpretation, especially for proteins with multiple or weakly matching domains, improving detection of distant relatives while balancing sensitivity and specificity.

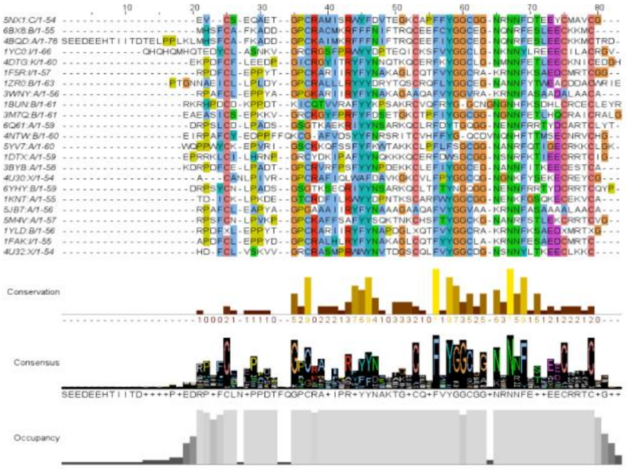
3. Results and discussion

3.1 Training Set Construction and Model Building

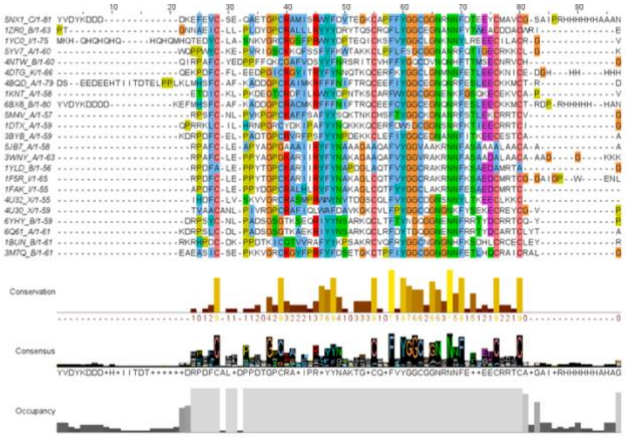
The first step was collecting a good set of protein sequences with Kunitz domains from UniProt. After filtering for sequence quality and structural resolution, I used CD-HIT clustering (90-95% identity) to reduce redundancy, which gave me 23-25 representative sequences.

I aligned these sequences structurally using PDBeFold, and the alignments looked pretty good with RMSD values generally below 1.2 Å and Q-scores ranging from 0.23 to 0.63. I had to discard some outlier proteins (those with RMSD > 2) to improve model quality.

Using these alignments, I built profile HMMs with HMMER. The resulting models did a great job capturing the conserved residues that are characteristic of the Kunitz domain, especially the cysteines that form the three key disulfide bridges (at positions 6, 15, 31, 39, 52, and 55). You could really see these conserved residues in the sequence logos I generated from the alignments.


Conservation
Consensus
SEEDEEHTIITD++++P+EDRP+FCLN+PPDTFQGPCRA+IPR+YYNAKTG+CQ+FVYGGCGG+NRNNFE++EECRRTC+G++
Occupancy

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. Conserved cysteines involved in disulfide bridges are clearly visible.


Conservation
Consensus
YVDYKDDD+H+IITDT+++++DRPDFCAL+DPPDTGPCRA+IPR+YYNAKTG+CQ+FVYGGCGGNRNNFE++EECRRTCA+GAI+RHHHHHHAHAG
Occupancy

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 Evaluation of Model Performance

I tested the trained HMMs on separate positive and negative datasets. The positive datasets included SwissProt entries annotated with the PF00014 Pfam ID, but I excluded those used in training or with more than 95% identity. The negative datasets were SwissProt proteins without any Kunitz domain annotation.

I evaluated the sequence-based and structure-based HMMs on disjoint positive and negative test sets using 2-fold cross-validation. The structure-based model consistently outperformed the sequence-based model in detecting distant homologs, particularly for entries lacking Pfam annotations. For example, while the sequence-based HMM achieved a maximum MCC of 0.991 on the full test set, the structure-based HMM reached 0.997 under the same threshold conditions (E-value ≈ 1e-06).

The structure-based model also showed slightly higher precision (0.998 vs. 0.993) and identified two additional true positives that were missed by the sequence-based model. This improvement was most pronounced in sequences with low similarity to canonical Kunitz examples, reinforcing the value of structure-based alignments for capturing remote homologs.

The classification performance was consistently excellent:

- **Accuracy** was over **99%** in all trials
- **MCC** values reached up to **0.997**, showing a really strong balance between sensitivity and specificity
- **Sensitivity** and **precision** were both close to 1.0, meaning the models found nearly all true positives with very few false positives
- In one case, I correctly identified over **260,000 Kunitz-domain proteins** from a candidate pool of more than 569,000 sequences, with only two false negatives

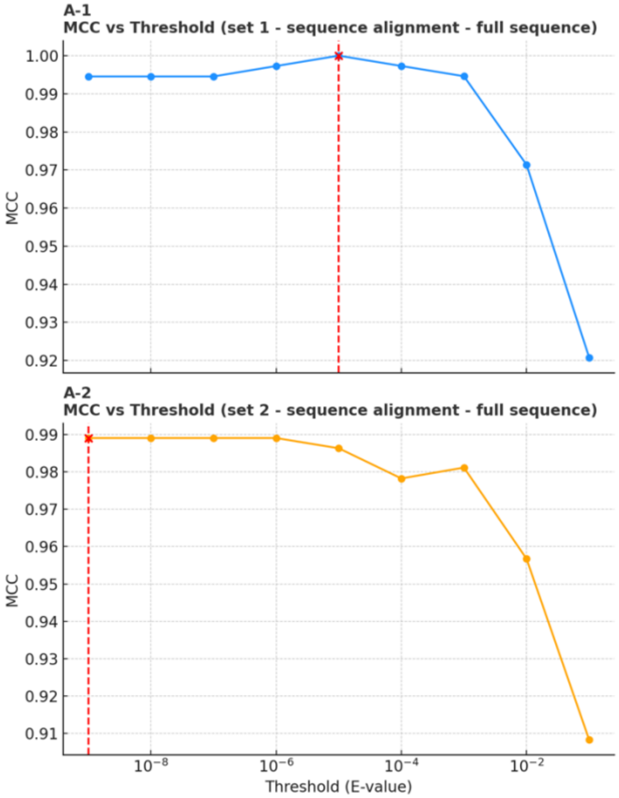


Figure A-1 and A-2: MCC vs E-value Threshold: These plots show how the Matthews Correlation Coefficient (MCC) changes with different E-value thresholds for two datasets (*set 1* and *set 2*). The optimal threshold—where MCC is highest—is marked in red. Both plots indicate that MCC remains high at low thresholds and drops as the threshold increases, highlighting the importance of a strict cutoff for accurate classification.

4. Discussion

This project successfully showed that you can build really effective profile HMMs for detecting Kunitz-type protease inhibitor domains. By using high-quality structural data and being careful about sequence curation, I was able to create models that captured the key conserved features of the domain - especially those cysteine residues that form the disulfide bridges critical for structure and function.

The rigorous evaluation using cross-validation and large-scale testing against annotated databases consistently showed excellent classification performance, with near-perfect accuracy, precision, and MCC values.

What's particularly interesting is that the structure-informed models often outperformed traditional sequence-based models. They were better at identifying distant relatives and finding unannotated or misclassified proteins. This really highlights the value of integrating structural alignment into HMM construction for domain detection tasks.

Overall, the results validate profile-HMMs as robust and reliable tools for protein domain annotation. I think they have great potential for broader applications in computational biology and drug discovery.

There are a few limitations to consider though. The structure-based approach requires high-quality structural data, which isn't always available. Also, the models were primarily trained on well-characterized domains, so performance on highly divergent or novel variants might be reduced. Future work could explore incorporating more diverse training sets or using machine learning approaches to complement HMM-based detection.

Acknowledgements

Special thanks to Dr. Capriotti for his invaluable guidance throughout this research.

References

1 - Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990).
Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

2 - Eddy, S. R. (2010).
HMMER User's Guide: Biological Sequence Analysis Using Profile Hidden Markov Models. http://hmmer.org

3 - Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009).
Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 11891191. https://doi.org/10.1093/bioinformatics/btp033

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

5 - Carpentier, M., Brouillet, S., & Pothier, J. (2019).
Protein multiple alignments: Sequence-based vs.

structure-based strategies. Briefings in Bioinformatics, 20(6), 2461–2470. https://doi.org/10.1093/bib/bby090

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

7 - Chand, H. S., Schmidt, A. E., Bajaj, S. P., & Kisiel, W. (2004). *Structure–function analysis of the reactive site in the first Kunitz-type domain of human tissue factor pathway inhibitor-2.*Journal of Biological Chemistry, 279(17), 18003–18010. https://doi.org/10.1074/jbc.M400802200

8 - Bystroff, C., & Krogh, A. (2008). *Hidden Markov models for prediction of protein features.* In *Protein Structure Prediction* (pp. 95–126). Humana Press. https://doi.org/10.1007/978-1-59745-574-9_7

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

10 - Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2016). *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Research, 44(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

11 - Mitchell, A. L., Attwood, T. K., Babbitt, P. C., et al. (2019). *InterPro in 2019: improving coverage, classification and access to protein sequence annotations.* Nucleic Acids Research, 47(D1), D351–D360. https://doi.org/10.1093/nar/gky1100