Bioinformatics

Development of a Profile Hidden Markov Model for the Kunitz-type Protease Inhibitor Domain

Department of Pharmacy and Biotechnology (FaBit), Alma Mater Studiorum – University of Bologna

Abstract

Motivation:

The Kunitz domain is a small, structurally conserved motif with key roles in protease inhibition and diverse biological functions across species. Accurate detection of this domain is essential for functional annotation and comparative studies.

Results:

This study presents a structure-informed Hidden Markov Model (HMM) for the identification of Kunitz domains. Starting from a curated dataset of 159 Kunitz-containing protein structures from the PDB, the authors performed structural alignments and derived a multiple sequence alignment to build a profile HMM using HMMER. The model was validated against positive and negative datasets extracted respectively from PDB and Swiss-Prot, with redundancy reduction and stringent filtering applied. Through cross-validation, the model achieved high accuracy in distinguishing true Kunitz domains, with Matthews Correlation Coefficients (MCC) consistently above 0.99 and minimal false positive/negative rates. Both full-sequence and domain-based scoring approaches were tested, yielding robust and transferable E-value thresholds for classification

Contact: valerio.piersanti@studio.unibo.it

Supplementary information: https://github.com/Valerio-Piersanti/Lab-project

1 Introduction

1.1 Kunitz domain

The Kunitz domain [1] (Pfam 00014) is a small, evolutionarily conserved protein domain commonly involved in the inhibition of serine proteases. A typical Kunitz domain weighs about 7 kDa [2] and consists of about 60 amino acids and adopts a compact α+β fold that is stabilized by three conserved disulfide bridges, arranged in a specific bonding pattern: C1–C6, C2–C4, and C3–C5. Among these, the C1–C6 and C3–C5 bridges are essential for maintaining the domain's native conformation, while the C2–C4 bridge contributes to stabilizing the protease-binding region, though it is not strictly necessary for inhibitory function. Structurally, the domain includes a distorted antiparallel β-sheet core, flanked by a short helical segment at the N-terminus and a longer α-helix near the C-terminus. The folding brings the termini into close proximity, enhancing the overall stability of the molecule. This stability, mainly due to its disulfide-rich architecture, makes the Kunitz domain a reliable scaffold for protease inhibition. Interestingly, some natural variants of the domain, such as those found in venomous animals, lack one of the three disulfide bridges but still retain both structural integrity and inhibitory activity. This shows that the Kunitz domain is not only highly stable but also functionally adaptable across different biological contexts [1]. Kunitz-type inhibitors exhibit alternative functions depending on the organism. In venomous invertebrates like scorpions and cone snails, they can act as both neurotoxins and protease inhibitors. In parasitic helminths, Kunitz inhibitors play a key role in protecting the parasite from the host's digestive proteases. In nematodes, proteins with Kunitz domains are involved in collagen biosynthesis, while some can trigger IgE-mediated allergic reactions. In blood-sucking arthropods, most Kunitz inhibitors function as anticoagulants, and several also serve as defense molecules against microbial pathogens [1].

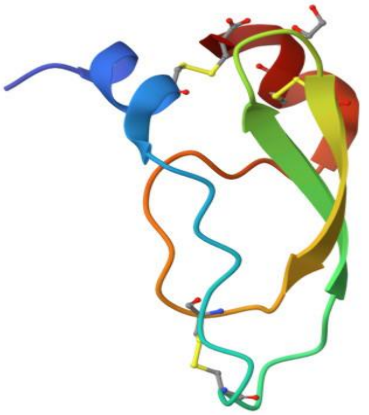

Figure 1Kunitz domain PDB DOI: https://doi.org/10.2210/pdb8AJ7/pdb

1.2 Hidden Markov Model

Hidden Markov Models (HMMs) are a type of unsupervised learning technique applied to sequential data. They are designed to model a Markov process in which a sequence of observed values {x1, ..., xT} is believed to arise from a series of hidden underlying states {s1, ..., sT}, which can be discrete. In an HMM with K states, the hidden states si are assumed to take values from the set {1, ..., K}. The model parameters are estimated by maximizing the likelihood of the complete data [3]. In this paper, we will develop a Hidden Markov Model (HMM) to be used for the prediction of the Kunitz domain by using dataset derived from databases online (like UniProtKB/Swiss-Prot).

2 Methods

2.1 Dataset creation

To develop a Hidden Markov Model (HMM), it is necessary to construct datasets for training the model to recognize the Kunitz domain. For this purpose, we use the Protein Data Bank (PDB) to retrieve all protein structures containing the Kunitz domain, applying the following selection criteria: Pfam annotation PF00014, data collection resolution ≤ 3.5 Å, and polymer entity sequence length between 45 and 80 amino acids. As a result, 159 proteins were obtained. The PDB was selected as the primary source for this study due to its comprehensive and publicly accessible repository of three-dimensional structural data for biological macromolecules, including proteins and nucleic acids. In the context of this project, the PDB provides experimentally resolved, high-quality structural data for proteins containing the Kunitz domain, which are essential for the development and training of a structure-based Hidden Markov Model. Subsequently, to reduce sequence redundancy and avoid bias during model training, CD-HIT is employed to cluster the sequences at a 90% identity threshold. A single representative sequence from each cluster is selected for use in subsequent analysis steps.

2.2 Multiple structure alignment

Using PDBeFold, a structural alignment was performed, yielding 25 matching protein structures. The alignment consisted of 28 aligned residues, with a root mean square deviation (RMSD) of 1.109 Å and a Q-score of 0.1429, indicating a moderate structural similarity across the aligned regions.

2.3 Hidden Markov Model

To prepare the data for Hidden Markov Model (HMM) construction, the aligned sequences were first processed to ensure compatibility with HMMER, a widely used tool for profile HMM construction. The initial step involved converting the multiple sequence alignment (MSA) from its original format into FASTA format. Following the preparation of the FASTA file, the next crucial step was the generation of the HMM model. Using the HMMER tool, the hmmbuild function was employed to construct the HMM from the aligned sequence data. This tool uses a probabilistic framework to create a model that captures the sequence conservation and variability within the aligned dataset.

2.4 Training sets

The positive and negative benchmark sets were derived from distinct sources. The positive set was extracted from the previously described dataset, sourced from the Protein Data Bank (PDB), containing proteins characterized by the Kunitz-type domain. In contrast, the negative set was compiled from Swiss-Prot, which includes all reviewed protein sequences in the database, and consists of sequences lacking the Kunitz domain.
To obtain the positive dataset, we first extracted the 25 identifiers of the proteins used in the structural alignment from the alignment file. Using these identifiers, we retrieved the corresponding full-length sequences from the comprehensive Kunitz protein dataset. These representative sequences were then compared against the entire Kunitz dataset using BLASTP. To avoid redundancy and potential bias in downstream analyses, we filtered out sequences showing ≥95% sequence identity and an alignment length of at least 50 residues, obtaining 364 sequences out of 397. This allowed us to retain only those sequences with lower similarity to the structurally aligned representatives for further analysis. Subsequently, we extracted from the dataset containing all proteins only those that did not include the Kunitz domain, thereby generating the negative dataset.

2.5 Testing and Evaluation of the HMM Model

The positive dataset was randomly shuffled and split into two subsets, each comprising 184 proteins. Similarly, the negative dataset was shuffled and divided into two equal subsets, each containing 286,286 proteins. At this stage, the hmmsearch command is executed to evaluate the proteins against the previously constructed model. This analysis is carried out across all four datasets. We employed the -Z 1000 flag in the hmmsearch command to normalize E-value calculations by simulating a database size of 1000 sequences. This adjustment was necessary to ensure comparability between positive and negative datasets, which differ substantially in the number of protein sequences they contain. Without this normalization, the E-values would be biased by dataset size, undermining the validity of performance comparisons. After performing

Training Set	Test Set	Best E-value Threshold	Accuracy	MCC	TPR	FPR
Set 1	Set 2	0.0001	0.99999	0.99184	0.99185	0.000005
Set 2	Set 1	1e-05	0.99999	0.99318	0.99726	0.000002
Set 1	Set 2	1e-06	0.99999	0.99318	1.00000	0.000000
Set 2	Set 1	1e-05	0.99999	0.99318	0.99726	0.000002

Table 1Model performance on training and test sets with optimal E-value thresholds, accuracy, MCC, TPR, and FPR values

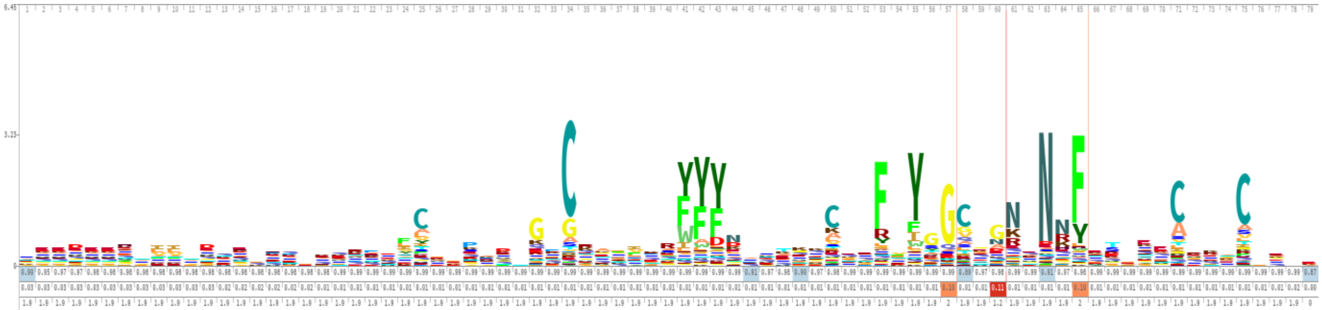

Figure 2 Amino acid conservation profile generated from the multiple structure-based alignment. The height of each stack indicates the level of sequence conservation at that position, with highly conserved residues reflecting key structural or functional features

hmmsearch, the IDs (labeled with 1 for kunitz and 0 for no kunitz) and E-values for both the sequences and the domains are extracted in .class files. The negative sets are then checked for any missing identifiers and updated with default E-values to ensure completeness in the dataset.

Finally, the positive and negative datasets are merged into two files to create a comprehensive training and testing set. This ensures the model is exposed to both classes, facilitating better classification performance by providing a balanced representation of positive and negative samples.

2.6 Results evaluation

To assess the robustness and transferability of the E-value threshold for classifying Kunitz domains, a cross-validation approach was employed. First, the optimal threshold was identified within one dataset by systematically testing a range of E-values (1e-1 a 1e-9) and selecting the one that maximized the Matthews Correlation Coefficient (MCC), a balanced metric that accounts for true and false positives and negatives.

This best-performing threshold was then applied to the second, independent dataset to evaluate its predictive performance outside of the training context. Key classification metrics such as precision, recall, and MCC were recalculated, and both false positives and false negatives were analyzed. The entire process was then repeated in reverse, using the second dataset to determine the optimal threshold and testing its applicability on the first dataset.

This reciprocal evaluation strategy provides insight into how well the threshold generalizes and helps identify potential overfitting to a specific dataset.

3 Results

Using the E-value of the full sequence as a score, the best threshold obtained from Set_1 was 0.0001. When applied to Set_2, this threshold yielded 183 true positives, 2 false positives, and 1 false negative, corresponding to a Matthews Correlation Coefficient (MCC) of 0.9919, a true positive rate (TPR) of 0.9892, and a positive predictive value (PPV) of 0.9946. The overall performance across both sets using this threshold resulted in 365 true positives, 3 false positives, and 3 false negatives, with an MCC of 0.9918, a TPR and PPV both equal to 0.9918. Conversely, using Set_2 to determine the optimal full-sequence threshold (1e-05), and applying it to Set_1 resulted in 181 true positives, 1 false positive, and 3 false negatives. This configuration achieved an MCC of 0.9891, a TPR of 0.9945, and a PPV of 0.9837. The corresponding overall results were 364 true positives, 1 false positive, and 4 false negatives, yielding an MCC of 0.9932, a TPR of 0.9973, and a PPV of 0.9891. When considering the best domain E-value instead of the full-sequence score, the optimal threshold from Set_1 was 1e-06. Applied to Set_2, this value resulted in 182 true positives, 0 false positives, and 2 false negatives, with an MCC of 0.9945, a TPR of 1.0, and a PPV of 0.9891. The corresponding overall evaluation showed 363 true positives, no false positives, and 5 false negatives, leading to an MCC of 0.9932, a TPR of 1.0, and a PPV of 0.9864. Finally, using the domain-based threshold of 1e-05 derived from Set_2 and applying it to Set_1 produced 181 true positives, 1 false positive, and 3 false negatives. The MCC obtained in this setting was 0.9891, with a TPR of 0.9945 and a PPV of 0.9837. The overall performance achieved 364 true positives, 1 false positive, and 4 false negatives, with an MCC of 0.9932, a TPR of 0.9973, and a PPV of 0.9891.

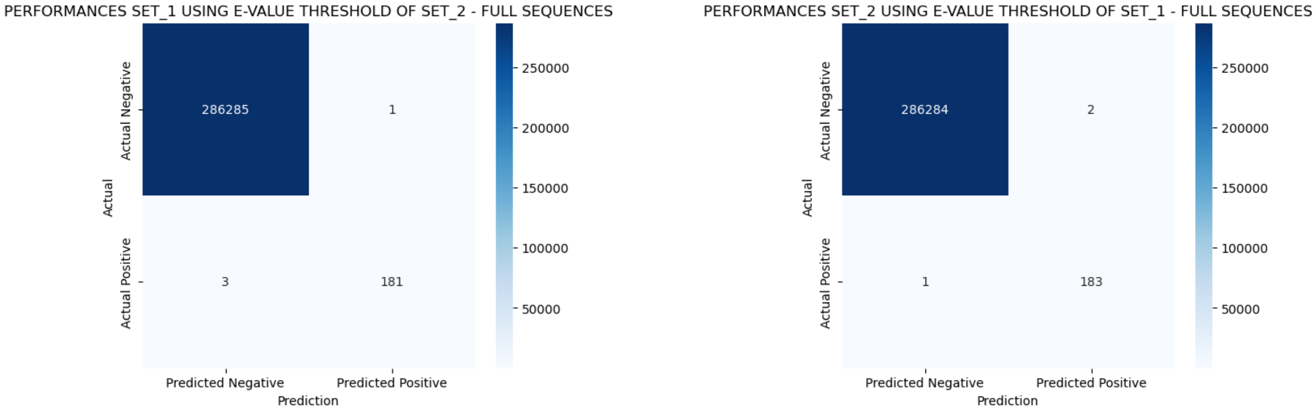

Figure 3 Confusion matrices showing true and false positive/negative counts for Kunitz domain prediction using optimal E-value thresholds. Left: results on Set_1 tested with threshold from Set_2; right: results on Set_2 tested with threshold from Set_1. The model

4 Discussion and conclusion

The results obtained from the evaluation of the predictive model demonstrate its strong overall performance in distinguishing between positive and negative classes. Across different e-value thresholds and datasets, the model consistently shows a high level of accuracy and robustness.

The Matthews Correlation Coefficient (MCC) values for both full sequence and single domain analyses were consistently above 0.99, indicating that the model has excellent predictive power. The high MCC values suggest that the model is capable of effectively identifying both true positives (TP) and true negatives (TN) while minimizing the occurrence of false positives (FP) and false negatives (FN). This is a crucial outcome for ensuring reliable predictions. Despite the strong overall performance, a small number of misclassifications were observed, particularly false negatives and false positives. These errors highlight the inherent challenge of optimizing thresholds for complex biological data, yet the overall performance of the model remains highly satisfactory. These misclassifications could potentially be minimized by further refinement of the thresholds or through additional model tuning.

In conclusion, the results demonstrate that the model is highly effective for the task at hand, providing reliable predictions across both full sequence and single domain analyses. The high MCC, TPR, and PPV values confirm its robustness, and the model appears well-suited for practical applications

References

[1] Shiwanthi Ranasinghe, Donald P. McManus, Structure and function of invertebrate Kunitz serine protease inhibitors, Developmental & Comparative Immunology, Volume 39, Issue 3, 2013, Pages 219-227, ISSN 0145-305X
https://doi.org/10.1016/j.dci.2012.10.005.

[2] Larissa Almeida Martins, Jan Kotál, Chaima Bensaoud, Jindřich Chmelař, Michail Kotsyfakis, Small protease inhibitors in tick saliva and salivary glands and their role in tick-host-pathogen interactions, Biochimica et Biophysica
Acta (BBA) - Proteins and Proteomics, Volume 1868, Issue 2, 2020, 140336, ISSN 1570-9639
https://doi.org/10.1016/j.bbapap.2019.140336.

[3] Meenakshi Khosla, Keith Jamison, Gia H. Ngo, Amy Kuceyeski, Mert R. Sabuncu, Machine learning in resting-state fMRI analysis, Magnetic Resonance Imaging, Volume 64, 2019, Pages 101-121, ISSN 0730-725X
https://doi.org/10.1016/j.mri.2019.05.031