Construction of a Hidden Markov Model (HMM) For the Kunitz Protease Inhibitor Domain

Rajab Ali

Department of Pharmacy and Biotechnology (Fabit), Alma Master Studiorum – University of Bologna
To Whom correspondence should be addressed

Abstract

The Kunitz domain is a structurally and functionally significant motif involved in protease inhibition across diverse biological systems. The study describes the creation of a Hidden Markov Model (HMM) designed to identify Kunitz type domains in protein sequences. The model constructed using a multiple sequence alignment derived from structural superposition of representative proteins, ensuring the incorporation of conserved structural features. The HMM was evaluated as a binary classifier on curated datasets from Uniport/SwissProt, demonstrating high accuracy and reliability through cross validation and extensive testing Performance metrics such as accuracy and the Matthews Correlation Coefficient (MCC) confirmed the model's robust classification capabilities, Additionally, this work underscores the advantages of integrating structural and sequence -based data to enhance domain recongnition while also revealing limitations in current annotation resources.

Contact: *rajab.ali2@studio.unibo.it*

1.Introduction

The Kunitz-type domain (Pfam: PF00014, InterPro: (IPR002223) is a Critical functional motif known for its role in inhibiting proteases, including serine cysteine and aspartic classes. Initially identified in the bovine pancreatic trypsin inhibitors (BPTI), this domain typically consists of approximately 60 amino acids and features a compact structure stabilized three disulfide bridges formed by six conserved cysteine residues. Its conserved fold includes two antiparallel β sheets and one or two helical regions, with structural integrity maintained by these disulfide bonds (C1-C6, C2-C4, AND C3-C5). These bonds not only stabilize the domains conformation but also facilitate interactions with target proteases via the protease binding loop.

Figure 11pdb_00002zjx Bovine pancreatic trypsin inhibitor (BPTI) containing only the [5,55] disulfide bond

Accurate prediction of the Kunitz domain in protein sequences is challenging due to structural and functional

variability across species. Hidden Markov Models (HMMs) offer a powerful computational approach to address this challenge. HMMs are statistical models that represents sequences of observations, assuming and underlying Markov process with hidden states. In protein domain analysis, HMMs model and predict domain presence by capturing position-specific conservation, gaps and insertions. Unlike simple sequence profiles, profile-HMMs incorporate probabilistic modeling to account for evolutionary variability.

The study presents method for developing an HMM tailored to Kunitz domain prediction. A multiple sequence alignments was generated from structural alignments of Kunitz containing proteins, serving as the foundation for the HMM profile constructed using the HMMER Pythons library. Prediction parameters were optimized to enhance accuracy, and the models performance was assessed using a Swiss -Prot dataset to evaluate its ability to distinguish Kunitz containing proteins from non -Kunitz sequences.

2.Materials and Methods

2.1 Data collecting and processing

High-quality protein structures annotated with the Kunitz domain were retrieved from the RCSB PDB, applying filters for resolution (≤3.OA), sequence length (50-80 amino acids), and absence of engineered mutations. Redundancy was reduced using CD-HIT with a 90% sequence identity threshold, yielding clusters of homologous proteins.

2.2 Structural Alignments

The refined FASTA file was curated to remove sequences with excessive length, insufficient length, or unstructured tails Representative structures from each cluster were selected for multiple structural alignment using PDBeFOLD, resulting in a final set of 23 distinct structures with associated metadata (PDB ID, sequence, chain ID and publication details).

2.3 Hidden Markov Model Construction (HMM)

The multiple sequence alignment was converted to FASTA Format and used to build the HMM profile via the HMMER 3.4 package in BioConda Model visualization was performed using Skylign, generating a sequence logo to illustrate residue conservation across alignment positions. The HMM was trained exclusively on cluster representatives to ensure non redundancy, with redundant proteins excluded during testing set creation.

2.4 Benchmark Data set Generation

Two datasets from UniProt/SwissProt were used for training and evalution:

Positive Set: 398 reviewed sequences with Kunitz domains, filtered to 368 sequences (95% similarity threshold, ≥50 amino acids) to exclude training -related sequences.

Negative Set: 573230 reviewed sequences lacking Kunitz domains

2.5 Model Testing

The datasets were shuffled and split into equal subsets (Positive: 184 protiens each , negative 286,286 protiens each). The hmmsearch command was executed with the -Z 1000 flag to normalize E-values, simulating a database size of 1000 sequences for fair comparison. Results were compiled into class files with missing identifiers in negative sets assigned default E- values. Positive and negative datasets were merged to create balanced training and testing sets.

2.6 Performance Evaluation

A Cross Validation approach identified the optimal E-value threshold (1e-1 to 1e-9) by maximizing the MCC . The threshold was validated on an independent dataset, with precision, recall and MCC recalculated. The process was repeated in reverse to assess generalizability. Performance was evaluated using full -sequence and best-domain E-values, yielding four rounds of assessment per subset. Metrics included accuracy ,MCC, true positive rate(TPR), and false positive rate (FPR).

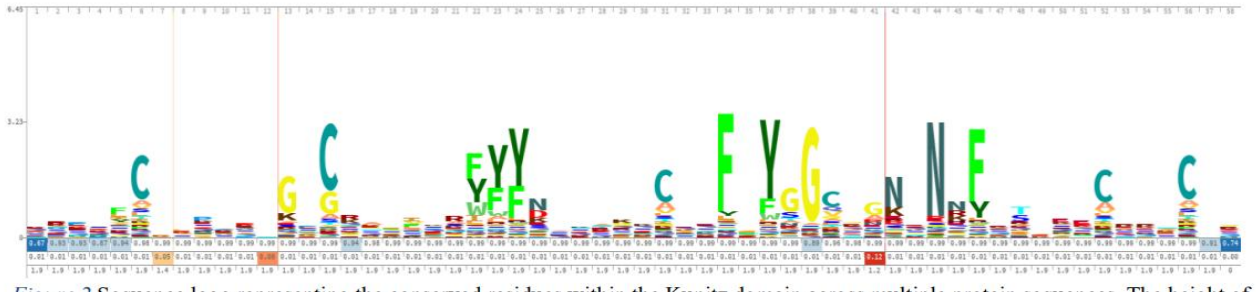

Figure 2Sequence logo representing the conserved residues within the Kunitz domain across multiple protein sequences. The height of

3.Results

The HMM exhibited exceptional performance in 2fold cross- validation, alternating between training and testing on subsets SET_1 and SET_2 key findings included.

Full sequence E-value mode (threshold: 1e-09): 181 true positives, 3 false negatives, 0 false positives,3 false negatives 0 false positives (MCC: 0.991, TPR 1.0, precision 0.984)

Best -domain E-value mode (threshold: 1e-06): identical performance to full sequence mode**.**

Reverse validation (SET-2 and SET-1): 2 false negatives, 0 false positives (MCC: 0.9955) , best -domain mode introduced 1 false positives (MCC :0.9943, Precision: 0.981).

The model achieved near -perfect sensitivity and specificity, with MCC values consistently >0.99.False negatives corresponded to weak homologs (E-values: 1-5e -05 to 0.011), while false positives were negligible.

Training	Testing	Best E-value threshold	Accuracy	MCC	TPR	FPR
SET_1	SET_2	1e-09 (Full sequence)	0.999990	0.9918	1.00e+00	0.00e+00
SET_1	SET_2	1e-06 (Best domain)	0.999990	0.9918	1.00e+00	0.00e+00
SET_2	SET_1	1e-06 (Full sequence)	0.999993	0.9945	1.00e+00	0.00e+00
SET_2	SET_1	1e-06 (Best domain)	0.999990	0.9918	1.00e+00	3.49e-06

Figure 3Performance metrics obtained with 2-fold cross-validation

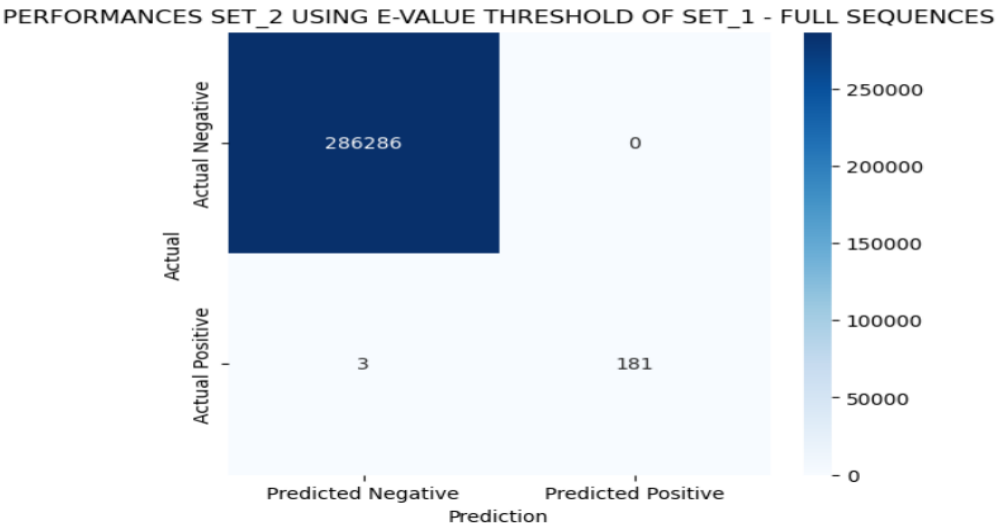

Figure 4Confusion matrix: training on SET_1, test on SET_2 – using full sequences E-value (best threshold= 1e-09)

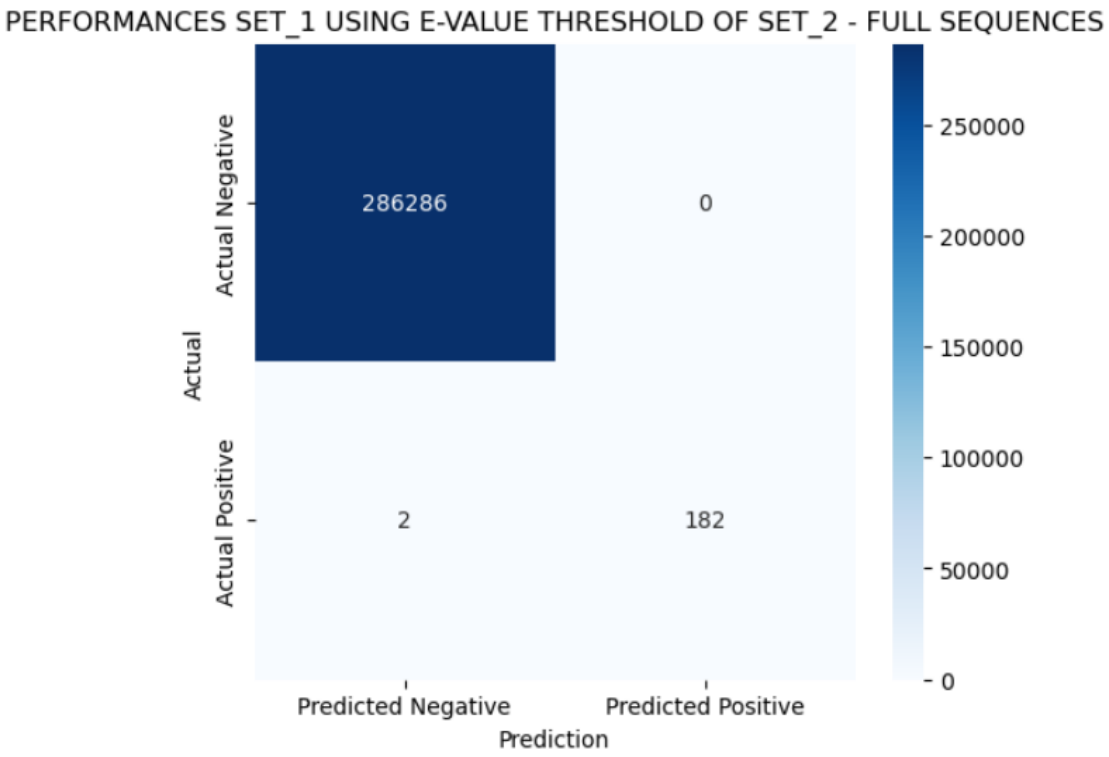

Figure 55 Confusion matrix: training on SET_2, test on SET_1, using full sequences E-value (best threshold=1e-06)

4. Discussion and Conclusion

The HMM demonstrated outstanding predictive power, with MCC, TPR AND precision across all evaluations. Minor misclassifications highlighted the challenge of threshold optimization but did not detract from overall reliability. The models robustness supports its use for Kunitz domain annotation and discovery, validating the structure derived HM approach. Further refinement could reduce marginal errors, but the current performance is well suited for practical applications in protein domain analysis.

References

1.S Ranasinghe and DP McManus. Structure and function of invertebrate kunitz serine protease inhibitors. Developmental & Comparative Immunology, 39(3):219–227, Mar 2013. Epub 2012 Nov 24.

2. K. Jablonowski. Hidden markov models for protein domain homology identification and analysis. In K. Machida and B. A. Liu, editors, SH2 Domains: Methods and Protocols, pages 47–58. Springer, New York, NY, 2017

3. Kunitz domain - an overview. ScienceDirect Topics. https://www.sciencedirect.com/topics/neuroscience/kunitz-domain.

4. 5 B.J. Yoon. Hidden markov models and their applications in biological sequence analysis. Current Genomics, 10(6):402–415, Sep 2009.