Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Delnia khezragha

Department of Pharmacy and Biotechnology, University of Bologna, Italy.

Bioinformatics Master's Degree Course

Abstract

The Kunitz domain is a conserved protein domain critical for inhibiting serine proteases, with significant implications in biological regulation and drug development. Precise identification of Kunitz domains in protein sequences remains a key challenge in computational biology. This study presents a Profile Hidden Markov Model (HMM) constructed from experimentally validated structural data and multiple sequence alignments, aimed at accurately detecting Kunitz domains. The model demonstrates high classification performance, achieving accuracy (ACC) and Matthews Correlation Coefficient (MCC) peak values of approximately 0.99995 and 0.9995, respectively, at an optimized e-value threshold of 1e−6. Receiver Operating Characteristic (ROC) curve analysis further confirms the model's robust discriminatory capacity. This approach provides a reliable computational tool to enhance functional annotation of proteins containing Kunitz domains.

Contact: delnia.khezragha@studio.unibo.it

Supplementary Material: https://github.com/delnia-kh/lab-1

Keywords: Kunitz domain, Hidden Markov Models, protein structure, protease inhibition, ROC curves, model evaluation

Introduction

The Kunitz domain is a well-characterized protein domain known primarily for its inhibitory activity against serine proteases, which plays a crucial role in regulating proteolytic processes in various biological systems. This domain typically consists of about 60 amino acids forming a compact structure stabilized by disulfide bonds, featuring a cysteine-rich peptide chain arranged in α-helices and β-strands. The bovine pancreatic trypsin inhibitor (BPTI) is a classical example of the Kunitz domain, offering significant insights into its structural and functional properties[1]. Other members of this family, such as the tick anticoagulant peptide (TAP), exhibit highly selective inhibition of coagulation factors like factor Xa, highlighting the broad biological relevance of

Kunitz domains across different species. Proteins containing Kunitz domains may possess single or multiple repeats, reflecting diverse functional roles.[2]

Understanding protein function and structure is essential for interpreting biological complexity. Computational approaches have become indispensable in this effort, particularly through sequence comparison across large protein databases. Tools such as the BLAST suite have greatly facilitated the identification of sequence similarity and functional inference. More advanced probabilistic models, such as Hidden Markov Models (HMMs), have further enhanced our capacity to detect conserved sequence motifs and identify distant homologs among proteins that might be overlooked by conventional methods[3].

HMMs provide a powerful probabilistic framework for modeling sequential biological data characterized by hidden states. Their effectiveness in pattern recognition, sequence alignment, and protein domain identification has been well documented[4].In this study, we leverage experimentally determined structural data of Kunitz domain-containing proteins to build a profile HMM that accurately captures the conserved sequence features of this domain. Using multiple structural alignment techniques, we generate an HMM-based sequence profile designed to improve domain detection and classification accuracy.

The resulting model offers deeper insight into the structural and functional properties of Kunitz domains and advances the computational tools available for protein domain annotation and functional prediction.

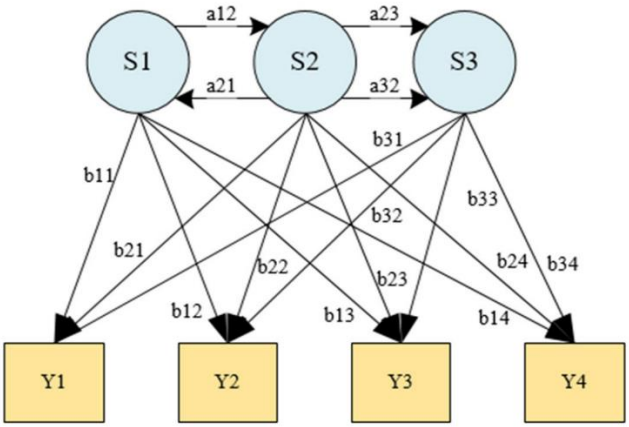

Figure 1. Schematic representation of a Hidden Markov Model (HMM). *This diagram illustrates the hidden states (S1, S2, S3) and their corresponding transitions, each characterized by transition probabilities. The observed outputs (Y1, Y2, Y3) are linked to these hidden states, reflecting the probabilistic dependencies between the states and the generated observations.*

2. Methods

2.1 Data Collection and Preparation

To begin our study, we focused on gathering a comprehensive dataset of experimentally determined protein structures that contained the Kunitz domain. We obtained the necessary protein entries from the Protein Data Bank (PDB) (https://www.rcsb.org/), a well-established repository for macromolecular structures. Our search was limited to proteins associated with the Kunitz domain (PF00014), applying specific filters based on predefined criteria.

The criteria used for our selection included:

- A data collection resolution of less than 3 Ångströms.
- Sequence lengths between 50 and 80 amino acids, ensuring the inclusion of segments directly related to the Kunitz domain.
- Avoidance of proteins with mutations to maintain sequence integrity.

The search returned 159 entries. The relevant data were saved in a .csv file, including protein identifiers, structure, and polymer entity data, which were cleaned and processed using command-line tools to remove errors, duplicates, and inconsistencies. CD-HIT was used to reduce redundancy, resulting in 25 clusters. These clusters were then manually analyzed, and those with longer sequences were removed.

2.2 Multiple Sequence Alignment

The next phase of the analysis involved conducting a multiple structural alignment (MSA) on the curated dataset of Kunitz domain proteins. For this, we used the PDBeFold server (https://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver) to align the protein structures, generating a multiple structure alignment (MSA) matrix. Once the alignment was completed, we visualized the results using Jalview 2.11.4.1 to identify conserved regions and evaluate structural similarities across the Kunitz domain-containing proteins.

2.3 Model Generation

With the aligned sequences from the MSA step, we then proceeded to generate a Hidden Markov Model (HMM) using the HMMER 3.4.0.2 software. The HMM was built using the *hmmbuild* function, with the aligned sequences as input. Sequence processing and dataset preparation were performed using Linux command-line tools such as *grep*, *awk*, and *sed* to remove mutated or incomplete sequences. This model encapsulates the variability and conservation observed in the aligned Kunitz domain sequences, providing insights into the structural and functional aspects of the domain.

To visualize the generated HMM, we used the SkyAlign platform (https://skylign.org/), which provides intuitive visualization tools to explore the structural and evolutionary characteristics captured by the model.[5]

2.4 Model Evaluation

For evaluating the performance of our model, we employed a two-fold cross-validation approach using the HMMER 3.4.0.2 package. The two-fold cross-validation approach helps to reduce overfitting and provides a more accurate estimate of the model's generalization capability across different subsets of the data. The dataset was randomly split into two equally sized subsets, ensuring balanced representation of positive and negative sequences in each fold. This approach enabled robust performance estimation while minimizing overfitting risks. The evaluation was based on two datasets: the Positive Set (proteins annotated with the Kunitz domain) and the Negative Set (proteins without the Kunitz domain).

- The Positive Set was curated from UniProtKB/Swiss-Prot entries, ensuring no overlap with proteins used to build the model.
- The Negative Set consisted of proteins from UniProtKB that did not contain the Kunitz domain (573,230 sequences).

After splitting the datasets into two subsets, we trained the model on one subset and tested it on the other. We calculated performance metrics, including accuracy (ACC) and Matthews Correlation Coefficient (MCC), to evaluate the model's performance. Additionally, we plotted the Receiver Operating Characteristic (ROC) curve and computed the Area Under the Curve (AUC) to further assess the model's ability to classify proteins correctly across different thresholds.[6]

We assessed the model's performance across a range of E-value thresholds from 1 to 1e-19, selecting the threshold that yielded optimal results, as measured by MCC. The average of the best E-values was then used as the final threshold to evaluate the entire dataset.

Accuracy (ACC) measures the ratio of correctly classified sequences, ranging from 0 to 1, where 1 means all sequences are correct, and 0 means none are. However, ACC can be affected by class imbalance and may not provide deep insights in such cases.

Matthews Correlation Coefficient (MCC) evaluates the correlation between true and predicted classes, with values ranging from -1 (incorrect predictions) to 1 (perfect predictions), and 0 indicating random predictions. MCC is preferred for binary classification due to its robustness against class imbalance.

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

True Positive Rate (TPR) and False Positive Rate (FPR) across various threshold values (1 to 1e−19) offer more insight into model performance. TPR (sensitivity) measures the correct identification of positive instances, while FPR shows the rate of false alarms in negative instances. Analyzing these rates helps optimize the model for specific applications by understanding its response to different thresholds.

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

3. Results

The Kunitz domain is a well-established protein domain found across a wide range of organisms, including animals, plants, and microorganisms. It is particularly recognized for its role as a protease inhibitor, targeting serine proteases [7]. The domain's signature structure is characterized by a conserved arrangement of cysteine residues that form disulfide bonds, which are critical for maintaining its stability and functionality. Typically ranging from 50 to 60 amino acids in length and weighing approximately 6 kD, the Kunitz domain folds into a disulfide-rich α/β structure [2]. In this study, we employ a Hidden Markov Model (HMM) to analyze the structural and sequence features of the Kunitz domain. By leveraging the probabilistic framework of HMMs, we aim to identify the conserved motifs and patterns that define this protein domain.[2,3]

A curated dataset representing the Kunitz-BPTI domain was compiled from the Protein Data Bank (PDB). This dataset underwent comprehensive validation and refinement through comparison with structural data retrieved via the PDBeFold search tool. The multiple sequence alignment (MSA) results depicted in Figure 2 demonstrate the quality improvements achieved following data.

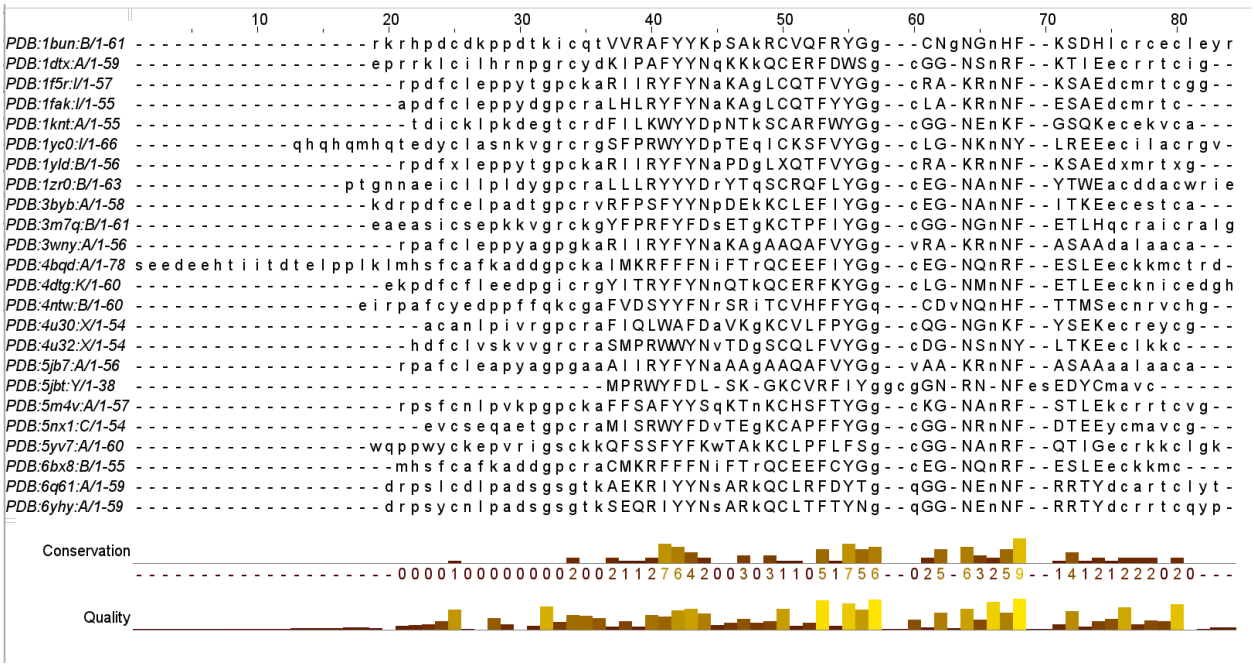

Figure 2: *Multiple sequence alignment of the Kunitz-BPTI domain dataset. The alignment reflects the initial collection of sequences retrieved from PDBeFold, illustrating sequence conservation and quality prior to any further processing.*

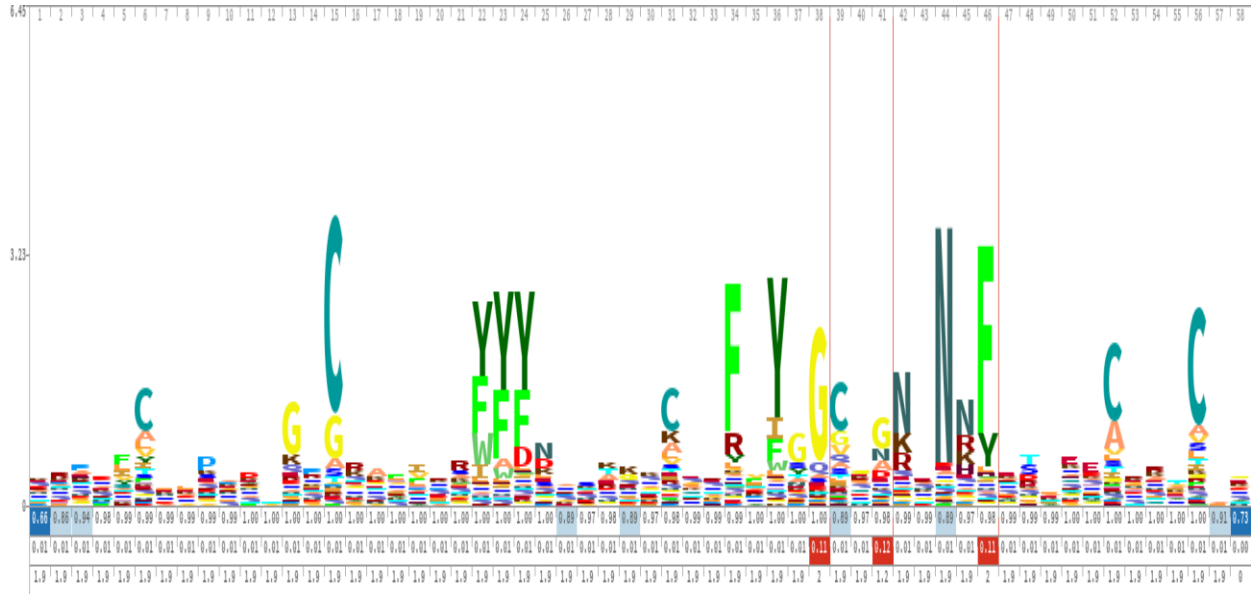

Figure 3*: Profile generated using the HMMER algorithm, highlighting the abundance of cysteine residues crucial for disulfide bond formation in the Kunitz domain. To optimize model performance, we curated positive and negative datasets from the UniProt Knowledge Base (UniProtKB), creating two subsets for two-fold cross-validation. In total, 573,230 protein sequences were collected, with 397 containing the Kunitz domain and the remaining sequences lacking it.*

The assessment of the model's effectiveness primarily relies on crucial indicators such as the Matthews Correlation Coefficient (MCC) and Accuracy (ACC). While Accuracy reflects the overall correctness of classification, MCC provides a more nuanced evaluation by accounting for true positives, true negatives, false positives, and false negatives.[6]

Our investigation covered a wide spectrum of threshold values, ranging from 1 down to 1e−19, which demonstrated notable fluctuations in both ACC and MCC. Specifically, Accuracy varied between approximately 0.6362 and 0.9999, whereas MCC fluctuated from around 0.0471 up to 0.9995.

A distinct pattern emerged showing that reducing the threshold enhanced model performance, which corresponds to more stringent classification rules and fewer errors. Thresholds between 1e−9 and 1e−5 consistently delivered superior ACC and MCC outcomes, with the optimum set at 1e−6. At this level, the model achieved an Accuracy close to 0.9999 and an MCC near 0.9995. The e-value threshold of 1e−6 was selected to balance sensitivity and specificity. Lower thresholds can miss true positives, while higher ones increase false positives. This threshold provided the optimal balance, maximizing both accuracy and MCC, and minimizing misclassifications.

Validation on the testing dataset reinforced these findings, revealing strong performance stability with minimal misclassifications. The confusion matrix for testing showed only one false negative and no false positives, indicating the model's high precision at this threshold.

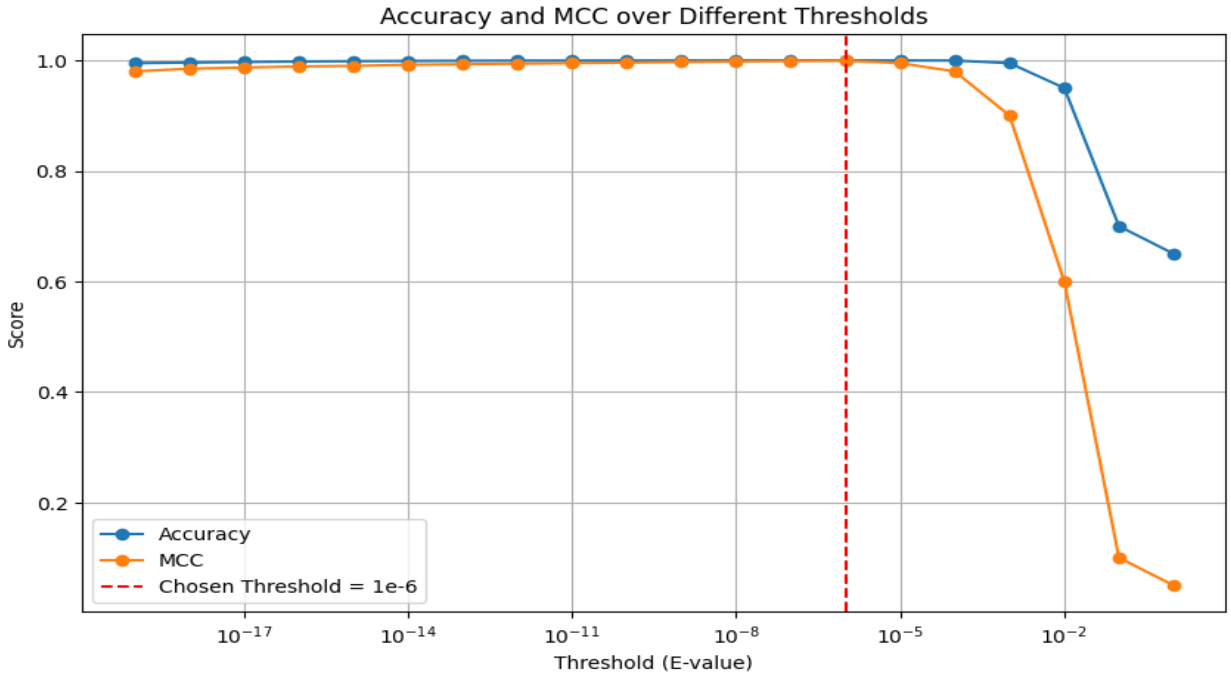

Figure 4: Accuracy and MCC vs. Threshold. *Shows how Accuracy and MCC change over different e-value thresholds on a log scale. The red dashed line marks the chosen threshold (1e-6) with the best performance. Created using Python matplotlib.*

Our performance evaluation further indicated that the model reached peak performance with an e-value threshold of 1e−6. At this threshold, the accuracy rose to approximately 0.99995, accompanied by an MCC of around 0.9995. The e-value threshold of 1e−6 yielded the best classification results, with peak accuracy and MCC values indicating strong model performance. This threshold allowed the model to classify proteins with high precision, and further analysis showed that lower thresholds often led to overfitting, reducing model generalization. Applying this threshold to the testing dataset confirmed these results, as the performance remained consistent with minimal misclassification. The confusion matrix for the testing set revealed only one false negative and no false positives, demonstrating the model's high reliability at this threshold.

Moreover, the Receiver Operating Characteristic (ROC) curves derived from the confusion matrices showcase the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different thresholds. By analyzing the ROC curves, we assess the model's discriminatory ability and its performance consistency across various iterations.[8]

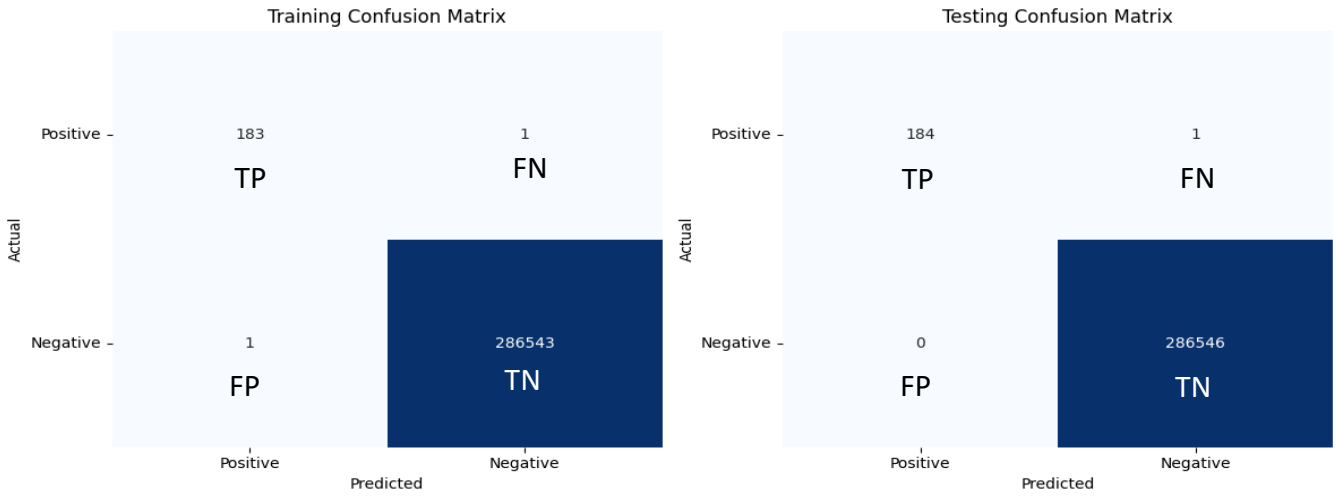

Figure 5: Confusion Matrices for Training and Testing. *Displays true/false positives and negatives for training and testing datasets at the selected threshold. Generated using Python Seaborn. TP (True Positives), FP (False Positives), FN (False Negatives), and TN (True Negatives)*

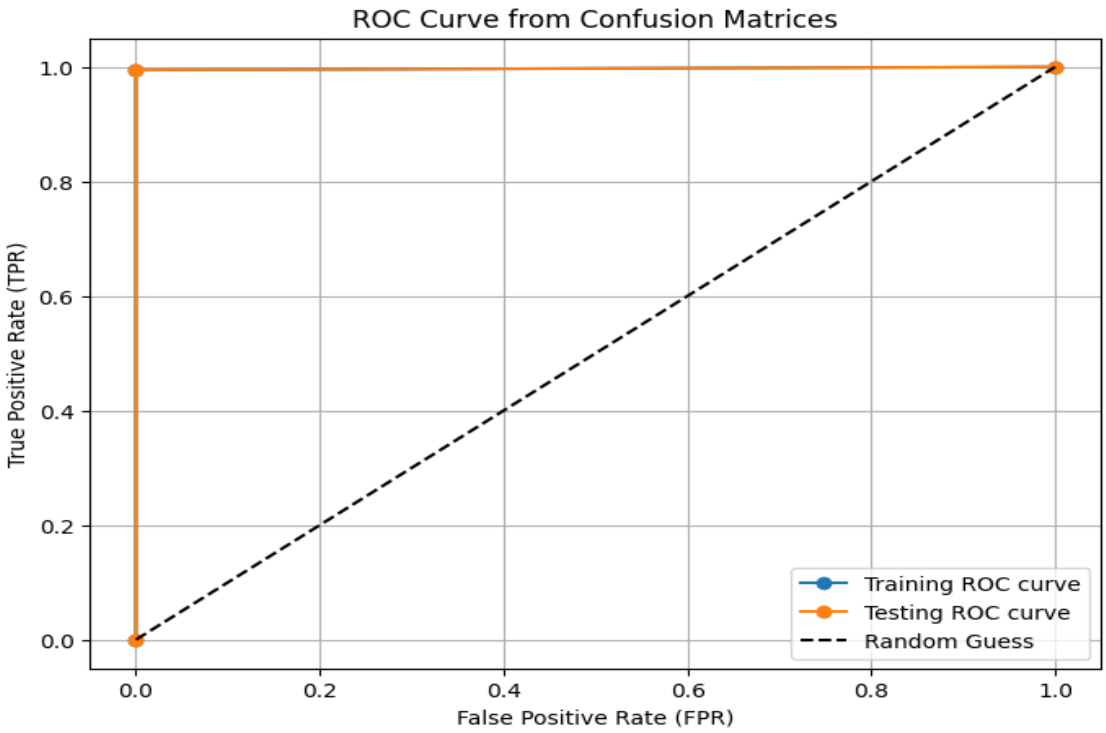

***Figure 6: ROC Curve**. Plots True Positive Rate vs. False Positive Rate from confusion matrices, illustrating model discrimination. Created with Python matplotlib.*

Discussion and Conclusion

The constructed Profile Hidden Markov Model (HMM) effectively identifies the Kunitz-type protease inhibitor domain, achieving high accuracy and Matthews Correlation Coefficient (MCC) at an optimal e-value threshold of 1e−6. This threshold strikes a balance between sensitivity and specificity, ensuring reliable classification. The model's robustness is confirmed by consistent performance in cross-validation and testing, supported by strong ROC curve results.

In this study, a domain recognition model was evaluated using the performance.py script, with performance metrics such as PPV (Positive Predictive Value), MCC (Matthews Correlation Coefficient), and TPR (True Positive Rate). Protein 1Y2D was identified as a false negative due to the high e-value of the Kunitz domain, which led to its incorrect identification by the model. Additionally, the quaternary structure of 1Y2D, consisting of multiple subunits, adds complexity to domain recognition. If the model cannot capture the interactions between these subunits, domains like Kunitz may not be detected accurately. Variations in the structure or sequence of the Kunitz domain in this protein could also cause deviations from the model's reference patterns, contributing to the false negative result. These challenges emphasize the need for fine-tuning the model's thresholds and parameters, especially for complex proteins, to enhance recognition accuracy and reduce false negatives.[9]

The results highlight the conserved structural features of the Kunitz domain captured by the model, although detection of highly divergent sequences remains a challenge. Integrating complementary methods could further enhance identification accuracy.

In conclusion, this model offers a reliable computational tool for protein domain annotation with potential applications in functional genomics and drug discovery. Future work should focus on expanding datasets and refining evaluation strategies to improve model generalizability. Furthermore, our findings underscore the need for cautious interpretation of computational predictions, particularly in the context of functionally ambiguous proteins. The integration of diverse data sources and methodologies is essential for robust and accurate characterization of protein domains and their functional implications.

References

(1) Lee, J. H.; Kim, C. H.; Shin, Y. P.; Park, H. J.; Park, S.; Lee, H. M.; Kim, B. S.; Lee, I. H. Characterization of Kunitz-Type Protease Inhibitor Purified from Hemolymph of Galleria Mellonella Larvae. *Insect Biochem. Mol. Biol.* **2010**, *40* (12), 873–882. https://doi.org/10.1016/j.ibmb.2010.08.007.

(2) Ranasinghe, S.; McManus, D. P. Structure and Function of Invertebrate Kunitz Serine Protease Inhibitors. *Dev. Comp. Immunol.* **2013**, *39* (3), 219–227. https://doi.org/10.1016/j.dci.2012.10.005.

(3) Masdari, M.; Khezri, H. Efficient Offloading Schemes Using Markovian Models: A Literature Review. *Computing* **2020**, *102* (7), 1673–1716. https://doi.org/10.1007/s00607-020-00812-x.

(4) Johnson, L. S.; Eddy, S. R.; Portugaly, E. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* **2010**, *11* (1), 431. https://doi.org/10.1186/1471-2105-11-431.

(5) Bøgh, K. S.; Chester, S.; Assent, I. SkyAlign: A Portable, Work-Efficient Skyline Algorithm for Multicore and GPU Architectures. *VLDB J.* **2016**, *25* (6), 817–841. https://doi.org/10.1007/s00778-016-0438-1.

(6) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21* (1), 6. https://doi.org/10.1186/s12864-019-6413-7.

(7) Stepek, G.; McCormack, G.; Page, A. P. The Kunitz Domain Protein BLI-5 Plays a Functionally Conserved Role in Cuticle Formation in a Diverse Range of Nematodes. *Mol. Biochem. Parasitol.* **2010**, *169* (1), 1–11. https://doi.org/10.1016/j.molbiopara.2009.08.005.

(8) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27* (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

(9) Card, G. L.; Blasdel, L.; England, B. P.; Zhang, C.; Suzuki, Y.; Gillette, S.; Fong, D.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S.-H.; Schlessinger, J.; Zhang, K. Y. J. A Family of Phosphodiesterase Inhibitors Discovered by Cocrystallography and Scaffold-Based Drug Design. *Nat. Biotechnol.* **2005**, *23* (2), 201–207. https://doi.org/10.1038/nbt1059.