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

Erfan Zohrabi

Bioinformatics Master's Degree, University of Bologna, Italy

Abstract

The Kunitz domain is a conserved protein domain involved in protease inhibition and various biological functions, often studied for drug discovery. This report outlines the construction of a Hidden Markov Model (HMM) for the Kunitz domain, facilitating the identification and classification of Kunitz domains within UniProtKB/SwissProt sequences and enabling the annotation of new proteins. The HMM revealed sequence characteristics and conserved residues of the domain, exhibiting high classification capabilities. Performance was evaluated using accuracy and Matthew's correlation coefficient, and results were visualized with ROC curves.

Supplementary Materials: GitHub Repository

Keywords: Kunitz domain, Hidden Markov Models, protein structure, protease inhibitor, computational prediction, model evaluation

Introduction

Proteins are fundamental molecular components involved in various biological processes, and understanding their structure and function is essential for deciphering the complexities of living organisms. This report examines the Kunitz protein domain, a structurally conserved domain present in diverse organisms, ranging from microbes to mammals. Known for its inhibitory activity against proteases, the Kunitz domain plays multiple biological roles, including involvement in inflammatory responses.

Kunitz-type domains are widespread in natural systems, functioning as serine protease inhibitors or as toxins in venomous animals. The Kunitz motif is a cysteine-rich peptide chain, approximately 60 amino acids in length, characterized by a compact fold with one or two α-helices and several β-strands forming an anti-parallel β-sheet, stabilized by disulfide bonds. In the Pfam database (v. 32.0), the Kunitz BPTI domain family (PF00014) includes over 30,000 sequences and 194 structures, organized into 2,475 potential architectures.

Bovine pancreatic trypsin inhibitor (BPTI), also known as basic protease inhibitor, is the prototypical example of the Kunitz domain family. However, this diverse family includes a wide range of members, such as snake venom basic protease, mammalian inter-alpha-trypsin inhibitors, tryptase inhibitor (a mast cell inhibitor of trypsin in rats), a domain in an alternatively spliced form of Alzheimer's amyloid beta-protein

(APP), domains at the C-termini of the alpha-1 and alpha-3 chains of type VI and type VII collagens, the tissue factor pathway inhibitor precursor (TFPI), and the Kunitz STI protease inhibitor found in legume seeds. These varied members within the Kunitz domain family illustrate the versatility and broad range of biological functions associated with this protein domain.

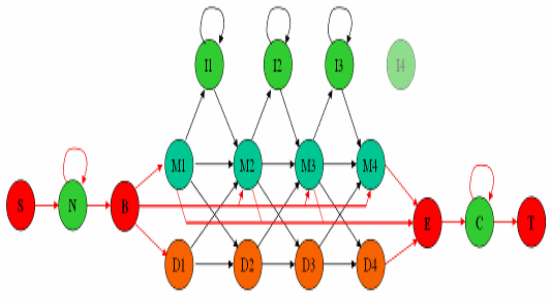

Fig1: A profile HMM is a statistical model that represents the consensus of a multiple sequence alignment of proteins in the same family. It has a repetitive structure of three states - Match (M), Insert (I), and Delete (D) - that correspond to the columns in the multiple alignment.

Profile hidden Markov models (profile-HMMs) have gained extensive use in applications such as protein classification and motif detection. Their popularity arises from their ability to efficiently and effectively represent sequence profiles. This report outlines the step-by-step process of constructing a Hidden Markov Model for the Kunitz protein domain, beginning with the acquisition of relevant sequence data from protein databases, followed by the construction of the model profile using the HMMER software package with the training sequences. This model is then applied to annotate Kunitz domains in the SwissProt database, identifying putative Kunitz domain-containing proteins and evaluating their significance through statistical analysis.

Materials and Methods

2.1 Model Data

The dataset used for constructing the HMM profile was curated through the following steps:

1. **Advanced Search on RCSB PDB**:
   - Filters applied: Pfam ID PF00014, resolution below 3 Å, and sequence length between 50 to 80 amino residues.
   - 151 sequences were retrieved.
2. **Grouping by Sequence Identity**:
   - Sequences were grouped by 95% sequence identity to select representative sequences, avoiding bias from overly similar structures.
   - 25 sequences were obtained.
3. **Inclusion of Auth Asym ID**:
   - To maintain consistency and avoid redundancy, Auth Asym ID was included in the Custom PDB Report, focusing on the chains of interest and selecting only representative chains from each unique structure.
4. **Multiple Structure Alignment**:
   - The dataset was loaded into a PDBeFold session for multiple structure alignment, resulting in an overall RMSD of 0.75 Å.
   - This alignment was used to build the model.

2.2 Positive and Negative Datasets

To train and test the model, two datasets were constructed from UniProt/SwissProt sequences:

- **Positive Set**:

    - Included all reviewed sequences containing the Kunitz-type domain PFam ID (PF00014).
    - 390 sequences were initially retrieved, with 374 remaining after removing those used to build the model to prevent redundancy and bias.
    - The ID Mapping function of UniProt was utilized, with 16 Uniprot IDs being retrieved and subsequently removed from the positive set using a script.
- **Negative Set**:
    - Included all reviewed sequences not containing the Kunitz-type domain PFam ID (PF00014), totaling 569,126 sequences.

2.3 HMM Building

The multiple sequence alignment from section 2.1 was used as input for constructing the HMM profile. This was achieved using the hmmbuild program of HMMER (v.3.3.2) with default settings. The WebLogo tool was employed to visualize the HMM by generating a sequence logo. This visualization revealed that residues involved in the first and third disulfide bridges (Cys6-Cys57, Cys31-Cys53) exhibited significantly higher information content compared to the second disulfide bridge (Cys15-Cys39). Phe34 and Tyr36 showed a high degree of conservation, primarily due to their role in stabilizing the reactive site structure through internal hydrophobic interactions.

2.4 Model Testing

Following the removal of proteins used for the model as detailed in section 2.2, the hmmsearch program of HMMER was executed against the entire dataset (positive and negative sets merged: 569,500 proteins) with parameters set to -Z 1, –noali, and -max to eliminate heuristics. This process identified 263,075 proteins with a Kunitz domain (E-value ≤ 1). The output was saved as a new dataset, retaining only the IDs and E-values. For proteins where the domain was not identified (306,425 proteins), an E-value of 100 was assigned using a bash command and then added to the previous set.

2.5 Performance Evaluation

The merged dataset was shuffled and divided into two subsets, which were used alternately as training and testing sets (subset1, subset2). Model performance was evaluated by running a script that took subset1, the list of positive protein IDs, and a threshold as input, producing the confusion matrix, accuracy (AC), Matthew's correlation coefficient (MCC), True Positive Rate (TPR), and False Positive Rate (FPR) values for thresholds ranging between 1 and 1e-20. These values were used in another script to plot the Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUC).

$$\mathrm{AC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

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

After identifying the threshold value ensuring optimal model performance, it was verified that the same results were achieved with subset2. This procedure was repeated by swapping subset1 with subset2. The average of the best E-values was then used as a new threshold to test the model against

the entire UniProt/SwissProt dataset (569,500 proteins).

$$TPR = TP / (TP + FN)$$

$$FPR = FP — (FP + TN)$$

Results

The following results were obtained when using subset1 as the training set and subset2 as the testing set. Performance evaluation indicated that the method performed optimally with an e-value threshold of 1e−08: the accuracy reached 0.999, and the MCC was 0.997. Within this range, the model correctly identified 99.4% of positives and all negatives. Using this threshold (1e−08) for the testing set resulted in an accuracy of 0.999 and an MCC of 0.995. Indeed, from the confusion matrix of the testing set, only one protein was incorrectly recognized in both positive and negative categories.

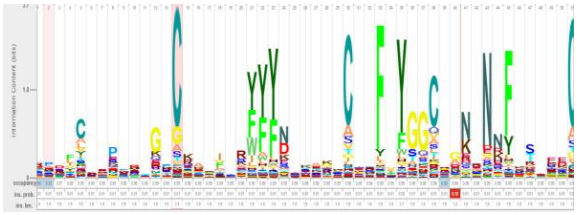

Figure 2. HMM Logo of the Kunitz-Type Domain: Generated by WebLogo, the sequence logo of the model illustrates the conserved residues within the Kunitz-type domain. The trimmed dataset used for this model includes all six conserved cysteine residues.

The ROC curve demonstrates near-perfect separability. The True Positive Rate (TPR) consistently reaches 1.0, indicating a high ability to correctly identify positive proteins, while the False Positive Rate (FPR) remains close to 0.0, indicating a low rate of misclassifying negative proteins. The AUC score is 1, underscoring the model's effectiveness in distinguishing between positive and negative proteins. When subset2 was used as the training set, the best threshold was in the range of 1e−09 to 1e−14. The testing set (subset1) achieved the same values for accuracy and MCC as described above.

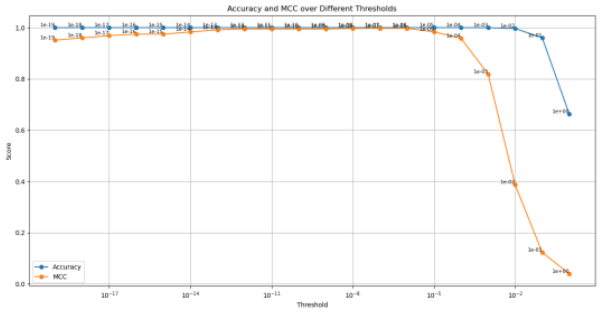

Figure 3: Accuracy and Matthews Correlation Coefficient (MCC) Across Various Thresholds

As verified, the two best thresholds are 1e-08 and 1e-09, resulting in an average of 5.5e-09. Using this threshold for the entire dataset showed an accuracy of 0.999 and an MCC of 0.997. The confusion matrix for the entire dataset indicates two false negatives.

Table1. Performance Metrics for Different Thresholds in Training and Testing Sets

Threshold	Training ACC	Training MCC	Testing ACC	Testing MCC
1e-07	0.9999824944682519	0.9860674955125075		
1e-08	**0.9999894966809512**	**0.9915952787926052**	**0.99999**	**0.99720**
1e-09	0.9999894966809512	0.991579492431331		
1e-10	0.9999894966809512	0.991579492431331		

Training		
	Predicted Positive	**Predicted Negative**
Actual Positive	177.0 (TP)	2.0 (FN)
Actual Negative	1.0 (FP)	285444.0 (TN)

Testing		
	Predicted Positive	**Predicted Negative**
Actual Positive	178.0 (TP)	1.0 (FN)
Actual Negative	0.0 (FP)	285447.0 (TN)

TP (True Positives), FP (False Positives), FN (False Negative), and TN (True Negative)

Table 2. Confusion Matrix for Training and Testing Sets at E-value of 1e−8.

Conclusions

Despite the promising outcomes, it is important to highlight the occurrence of two False Negatives (FN) in the final test set, specifically, the UniProtKB AC: O62247 and D3GGZ8 entries. In both cases, the Kunitz domain was not detected within the sequence because the e-value of the highest-scoring domain (2.4e-07 for O62247; 6.3e-06 for D3GGZ8) exceeded the threshold of 5.5e-09. The variation in structure and/or sequence within the domain of these two proteins could be a possible explanation. Both proteins have the same caution on their UniProt site pages, indicating that while they exhibit serine protease activity in vitro, it remains uncertain whether this activity is genuine due to the absence of serine protease characteristics in both proteins.

The objective of this study was to construct a Hidden Markov Model (HMM) specifically for the Kunitz BPTI domain using a dataset of structurally characterized proteins. Subsequently, the effectiveness of this model in identifying the presence of the domain in new seed sequences was assessed. The results demonstrate the success of the HMM-based approach in accurately predicting the occurrence of the domain in proteins known to possess it. The performance of the method is graphically represented by the Receiver Operating Characteristic (ROC) curve and the high value of the AUC. Optimal performance was observed within the range of e-value thresholds between 1e-08 and 1e-09, where the values of Accuracy and Matthews Correlation Coefficient (MCC) approached their maximum values, while the number of False Positives and False Negatives was negligibly low or absent. Consequently, it may be employed for functional annotation of unreviewed proteins.

References:

Salier JP. ”inter-alpha-trypsin inhibitor: emergence of a family within the kunitz-type protease inhibitor superfamily”. Trends Biochem. Sci., 15:435–439, 1990.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The jalview java alignment editor. Bioinformatics,

20(3), 426-427.

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), 17500-17507.

Tomii, K., Sawada, Y., & Honda, S. (2012). Convergent evolution in structural elements of proteins investigated using cross profile analysis. BMC bioinformatics, 13, 1-18.

Page AP. Stepek G, McCormack G. The kunitz domain protein bli-5 plays a functionally conserved role in cuticle formation in a diverse range of nematodes. Mol Biochem Parasitol., 169:1–11, 2010.

Chandonia JM Brenner SE. Crooks GE, Hon G. Weblogo: a sequence logo generator. Genome Res., 2004.

Mishra M. Aspects of the structural convergence and functional diversification of kunitz□domain inhibitors. J Mol Evol., 2020.

lvarez-Carreño, C., Gupta, R., Petrov, A. S., Williams, L. D., Alvarez-Carreño, C., Petrov, A. S., & Williams, L. D. (2022). Protein fold evolution by creative destruction. BioRxiv.