Teresa Gianni
Alma mater studiorum - University of Bologna
FaBit department, MSc Bioinformatics

Hidden Markov Model For Kunitz-type Protease Inhibitor Domain (Pfam: PF00014)

Abstract

The Kunitz protease inhibitor domain is a conserved structural motif found across multiple species and involved in a wide range of biological processes. As a result, there is increasing interest in developing computational models capable of accurately identifying this domain within protein sequences. In this work, is presented a methodology for building a Hidden Markov Model (HMM) aimed at classifying proteins based on the presence or absence of the Kunitz domain. The model was initially trained on a curated set of proteins known to contain the domain and subsequently tested on a larger dataset comprising previously unseen sequences. The model's performance was evaluated using multiple metrics, demonstrating highly promising results and supporting its effectiveness in domain identification tasks.
Supplementary material can be found in GitHub Teresa Gianni

1 Introduction

1.1 Kunitz-type Domain

Kunitz type proteins are a family of serine protease[1] inhibitors. The Kunitz motif is constituted by six conserved cysteine residues forming three disulphide bonds. The peptide chain under consideration typically contains approximately 60 amino acid residues. Its secondary structure includes a central α-helix and a small β-sheet, which is generally composed of two antiparallel β-strands stabilized by the disulphide bonds [1]. This motif has been observed in a variety of organisms, including vertebrates, invertebrates and certain plant species [2]. For instance, Kunitz inhibitors present in blood-sucking arthropods function as anticoagulants, while others act as a defence against microbial pathogens [3]. With regards to the vertebrates, Bovine Pancreatic Trypsin Inhibitor (BPTI) represents the prototype of this family of proteins and was the first Kunitz-type protease inhibitor to be described. Kunitz inhibitors are thus designated as BPTI-like proteins. They belong to the I2 family of peptidase inhibitors, which may contain one or multiple repetitions of the domain [4]. Some Kunitz proteases can act as ion channel blockers, are known as Kunitz-type toxins (KTT) and are frequent components of the venoms from poisonous animals. However, as previously stated, the primary function of this domain is the inhibition of serine proteases. This function is crucial in biological processes where the protease activity needs to be taken under control such as digestion, coagulation and fibrinolysis, inflammation, cell signaling and neuroprotection. It acts as a competitive inhibitor, firmly binding to the active site of serine proteases and thereby obstructing substrate access. This is possible thanks to the presence of a loop that mimics a natural substrate peptide. In particular, the reactive site P1—P1' peptide bond is located in the most exposed region of the protease binding loop. A significant proportion of the enzyme–inhibitor interface is characterised by the P1 residue, which penetrates deeply into the S1 specificity binding pocket of the protease. The integration of the P1 side chain inside the S1 binding pocket of the enzyme plays a major role in the energetics of the recognition. When the protease binds to the inhibitor, it begins to cleave the P1-P1' bond, but due to the rigid structure of the Kunitz domain (stabilised by disulfide bonds), the cleaved bond reforms rapidly, preventing the enzyme from completing the reaction. This binding is highly specific and high affinity. The result is a very stable enzyme-inhibitor complex in which the inhibitor effectively "freezes" the protease in an inactive state. This interaction is frequently reversible, but with a notably slow dissociation rate, resulting in prolonged inhibition. A classifcation of Kunitz inhibitors based on their inhibition profle can help in designing specifc inhibitors against proteases of various mechanistic classes by making defned alterations in the 60 amino acid scafold of the Kunitz inhibitor. This has been possible because residues in the loops can be substituted or grafted into desired ones without destabilizing the basic structural framework. Kunitz domain scafold is also present in human proteins and thus may possess very low immunogenic potential when used in therapeutics [5].

1.2 Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a statistical model used to represent systems that transition between a series of hidden (unobservable) states over time. Each hidden state generates observable outputs (emissions) based on certain probabilities. The key assumptions of an HMM is that to the transition between states applies a markov property, meaning that the future state depends only on the current state, not on the sequence of events that preceded it [6] [7]. In the specific case of this project, a HMM

[1] Enzymes that use a serine residue in their active site to cleave peptide bonds.

has been generated for solving a problem of binary classification of proteins as containing or not the Kunitz domain. Therefore, it is possible to interpret the Markov states as positions of a Multiple Sequence Alignment generated from a Multiple Structural Alignment (MSA) of protein used for training the model to recognise the subjected domain. Each state models the statistical profile of amino acid occurrences at that position within the Kunitz domain.

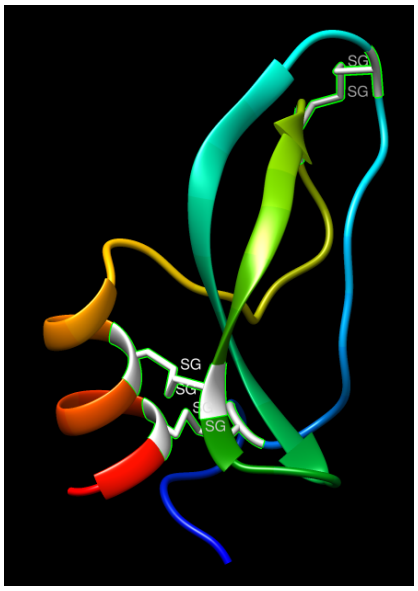


Figure 1: *Graphic representation of the Kunitz domain with ribbons shown in a rainbow color scale and the characteristic disulphide bonds shown in white. Image made with the software Chimera.*

2 Materials and Methods

2.1 Data Collection

The preliminary step in this procedure is to extract the following data from the public database, UniProt:

- The sequences of human proteins containing the Kunitz domain (18 sequences)
- The sequences of not-human proteins containing (377 sequences)
- The sequences of all the proteins presents in SwissProt database for testing the model on unseen data.

Subsequently, a custom report was generated, encompassing protein structures from the Protein Data Bank (PDB) public database that contain the kunitz domain. In particular, the information presents in the report were: auth asym ID, sequence, resolution, entity ID, annotation ID. These structures (in total 158) were selected based on criteria including a data collection resolution of 3.5 or less, and a sequence length ranging from 45 to 80 amino acids, which is approximately the length of the Kunitz domain. The objective of this operation was to determine whether the results were adequate for performing the analysis and, eventually, to identify the data that required cleaning.

2.2 Data Cleaning

From the custom report mentioned in the paragraph 2.1 it is evident that certain of these sequences exhibit a high degree of similarity, thereby resulting in a redundancy issue among the data. The CD-HIT algorithm has been utilised to facilitate the resolution of the issue. This algorithm allows for the clustering of sequences based on an identity threshold of 0.9 (90% identity), and the selection of a representative sequence for each cluster, reducing redundancy. From this procedure resulted 25 non redundant sequences which were than subjected to a preliminary Multiple Sequence Alignment with Muscle. This analysis revealed that the sequence corresponding to protein 2ODY was too long in comparison to the other sequences. In contrast, the sequence of protein 5JBT was found to be considerably shorter; therefore, they have been excluded from further consideration. Consequently, the final file contained 23 sequences.

2.3 Multiple Structural Alignment

This operation consisted in comparing the 3D structures of the 23 proteins obtained in the previous steps. The Multiple Structural Alignment (MSA) has been performed utilising the tool PDBeFold that receives a list of PDB protein identifiers as an input, and provides statistical values as an output. In this analysis, particular attention was paid to RMSD (*Root Mean Square Deviation*)[2], which was found to be less than 1Åin all cases. This indicates a high degree of similarity for each alignment. Once the MSA has been obtained and its suitability for further analysis has been confirmed, it is possible to generate a Multiple Sequence Alignment (shown in Figure 2) derived from the Multiple Structure Alignment. Such data have been stored in a *fasta* format and moved into an empty file (in the format *.ali) previously created.

2.4 Hidden Markov Model Build

At this point it is possible to build a Hidden Markov Model (HMM) from the the Multiple Structure Alignment file created in 2.3. This can be accomplished by utilising the HMMER software which can be installed in bash . The model's output indicates 58 match states among 84 positions. In order to facilitate the interpretation of the output of the HMM, it is useful to upload it to Skylign which will generate a logo (shown in Figure 3) of the model and allow the most conserved sites to be recognised since the height of each letter represents the relative frequency of the corresponding amino acid at that position.

2.5 Database Creation

The objective of this step is to create a *benchmark set*[3] to evaluate the performance of the model created at point 2.4. In this particular instance, such set consists of the entire SwissProt database of proteins that have already been annotated as either having or not having the kunitz domain. The set's classification will than be compared to the model's classification. Since the model was created with the 23 PDB sequences of the Kunitz domain, it is

[2]Measure in Ångström (Å) used to quantify the average distance between corresponding atoms of two superimposed molecular structures, typically proteins or nucleic acids.

[3]A set of which the classification is alredy known

crucial to eliminate the sequences that will exhibit high similarity among the entire set of Uniprot 2.1 Kunitz domain sequences (395 in total). This procedure involves establishing a BLAST database encompassing all Uniprot sequences (human and non-human) containing the Kunitz domain. Subsequently, the 23 proteins employed to construct the model were queried against this database, with a sequence identity of 95% across a minimum of 50 amino acids in the sequences being used as a filtering criterion. This comparison revealed 29 sequences with elevated levels of similarity that could affect the model's evaluation. These sequences are thus removed from the file with all the UniProt Kunitz sequences through the use of a python script that takes as input an IDs list and a fasta format file and gives as output the sequences relatives to the IDs. The sequences remaining after this operation will become the positive set (class 1) for the training containing 366 sequences. The sequences of the positive set created, should than be removed from the entire SwissProt database that will thus constitute the negative set (class 0) of the training with 572.864 sequences.

2.6 Model testing

In this phase, both the positive and the negative sets were randomly shuffled and divided into two subsets, resulting in four sets (*Pos 1, Pos 2, Negs 1, Negs 2*) containing respectively 183 (Pos) and 286.432 (Negs) sequences. The four sets have been submitted to a match with the HMM created at 2.4 utilising HMM search. Given the considerable disparity in dimensions between the sets, it is essential to establish a fixed number of sequences taken in consideration for the calculation of the *E-value*[4]. The output obtained from the match was converted into a tabular format including the protein ID and two E-values: one relative to the whole sequence and the other relative to the best domain matching. In view of the fact that the number of matches in the negative sets was found to be insufficient, a manual augmentation of solutions was necessary.

2.7 Performance Evaluation

For this procedure has been created two testing sets (Set 1, Set 2) by merging negative and positive sets of the steps before. For evaluating the performance has been used a Python script previously created that computed the following statistical measures on the set given:

- Build a **Confusion Matrix**: a table used to compare the predictions made from the model and the actual outcomes
- **Accuracy**

$$Q^2 = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Matthews Correlation Coefficient (MCC)**

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{2}$$

- **True Positive Rate (TPR)**

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

- **False Positive Rate (FPR)**

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

- **Precision**

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

This script takes as input one of the testing sets and an E-value as a threshold for discriminating among the classes (0 or 1). Therefore, the model's performance has been tested with different E-values ranging from $1e^{-1}$ to $1e^{-12}$ in order to choose the optimal E-value threshold. The final step entailed the execution of the performance on a newly constructed set, which was formed by the union of Sets 1 and 2. This was undertaken to ascertain the efficacy of the model when confronted with an extensively larger dataset, and to conduct a comparative analysis of the outcomes.

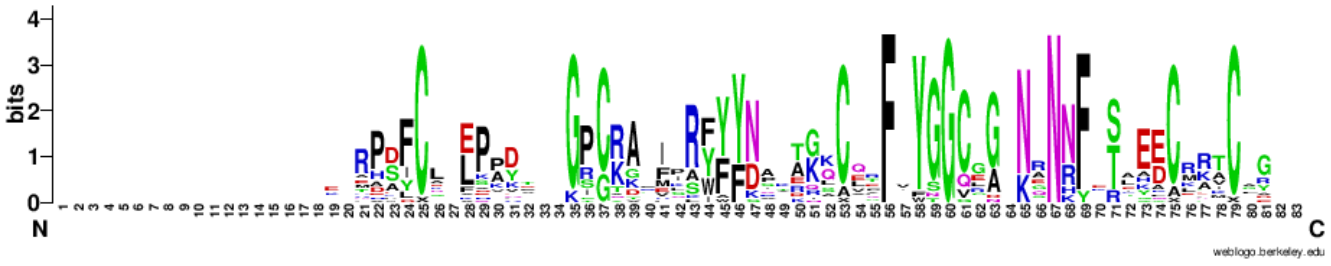


Figure 2: *WebLogo representing the Multiple Sequence Alignment obtained from the Multiple Structural Alignment*

[4]Statistical metric that represents the number of alignments with a given score that it can expect to find by chance in a database of a particular size. It quantifies the likelihood that a match between sequences is due to random chance rather than biological relevance.

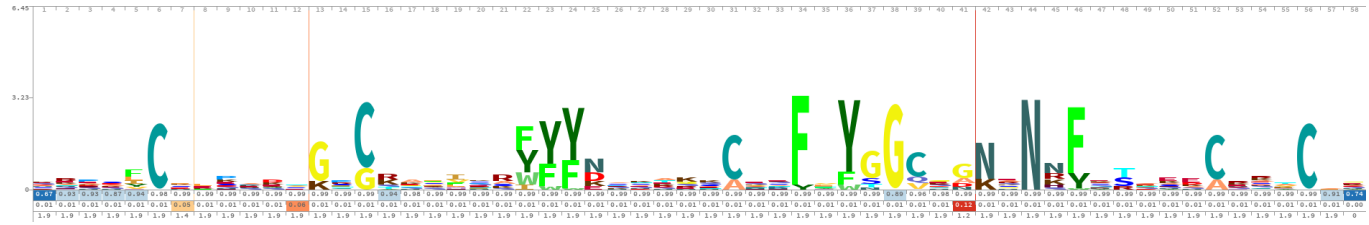


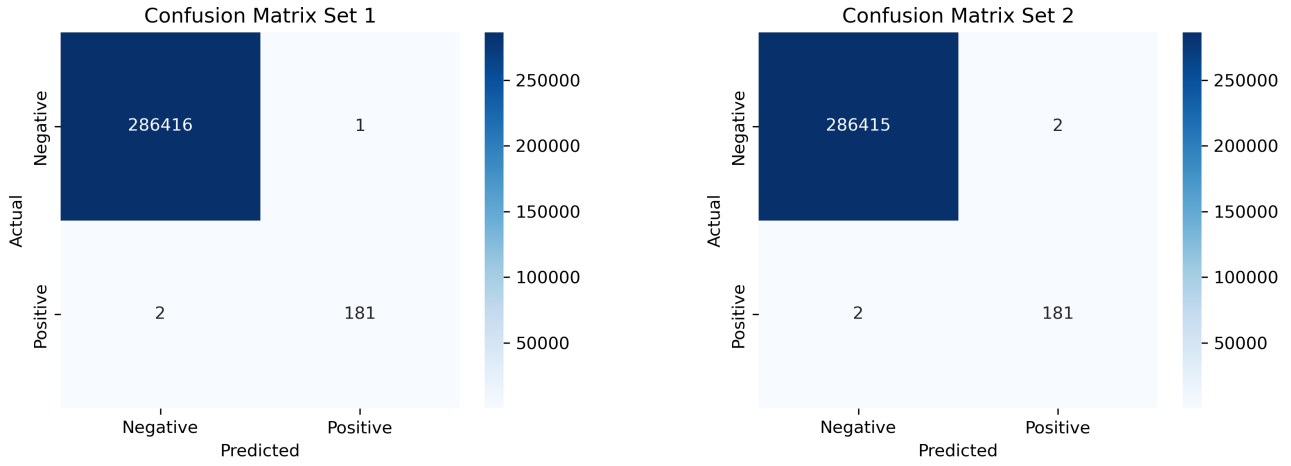
Figure 3: *Skylign logo that shows the training sequences of the Hidden Markov Model in terms of relative frequency*

3 Results

As illustrated in Table 1, each set demonstrated a remarkably high level of performance, with an accuracy of 0.999 and a precision of 0.989. There are only minor discrepancies in the rates of predictions (TPR and FPR) and MCC. However, it is worth noting that each value is extremely close to 1, which indicates an almost perfect classification. The reliability of the model is further supported by the confusion matrices presented in Figure 4, which demonstrate that the model made no more than four erroneous predictions out of over 280,000. A metric used for representing the trade-off between TPR and FPR is the *Receiver Operating Characteristic (ROC)* curve, which, in accordance with previous results, bows at the upper corner of the plot, thereby indicating an extremely high TPR and a low FPR. The ROC curve is frequently associated with the *Area Under the ROC Curve (AUC)*[5]Figure 6 algorithm, which, in this instance, is equivalent to 1, denoting optimal performance. On the other hand, the behaviour of MCC across various thresholds is represented in Figure 5 in which the stabilisation at an E-value of 1e-5 signifies the model's capacity to maintain high performance under increasingly stringent conditions.

Set	Best E-value threshold	Accuracy	MCC	TPR	FPR	Precision
1	10^{-6}	0.999	0.992	0.995	3.49×10^{-6}	0.989
2	10^{-6}	0.999	0.989	0.989	6.98×10^{-6}	0.989
1+2	10^{-6}	0.999	0.990	0.992	5.23×10^{-6}	0.989

Table 1: *Table representing the results of the performance analysis computed for the Hidden Markov Model*



(a) **(b)**

Figure 4: *Confusion matrices obtained from the performance analysis that indicate the number of True Positive (TP), False Negatives (FN), True Negatives (TN) and False positives (FP) model's predictions for both set 1 (Figure 4a) and set 2 (Figure 4b).*

4 Discussion and Conclusions

The extensive performance evaluations conducted provide substantial evidence that an Hidden Markov Model (HMM) trained on a Multiple Structural Alignment is a highly effective solution for the binary classification of the Kunitz domain's detection in protein sequences. This model has exhibited minimal variance in the performance of two independent sets and a substantially larger dataset, despite the considerable class imbalance. This demonstrates the model's high scalability, which makes it suitable for large-scale studies and adaptable to particular demands. The reliability of its high performance is evidenced in Figure 5 since it has reached a plateau over more stringent thresholds. Nevertheless, an investigation into the UniProt database revealed that the false negatives predicted by the model, corresponding to the IDs D3GGZ8, A0A1Q1NL17 and O62247, actually contain the Kunitz

[5]Summarizes the ROC curve's performance into a single value between 0 and 1.

domain, thus the model made an erroneous prediction. However, it is interesting to note that one protein (corresponding to the ID Q8WPG5) does not contain the Kunitz domain, denoting an error in the annotation of the Pfam database (from which testing data have been extracted). This can turn out to be an incredibly positive outcome, as it has the potential to compensate for erroneous or incomplete annotations. In conclusion, the results obtained from this project revealed an extremely powerful tool for binary classification problems. In future, the methodology could be applied to other datasets, or it could be exploited to develop analogous models with the aim of solving similar problems.

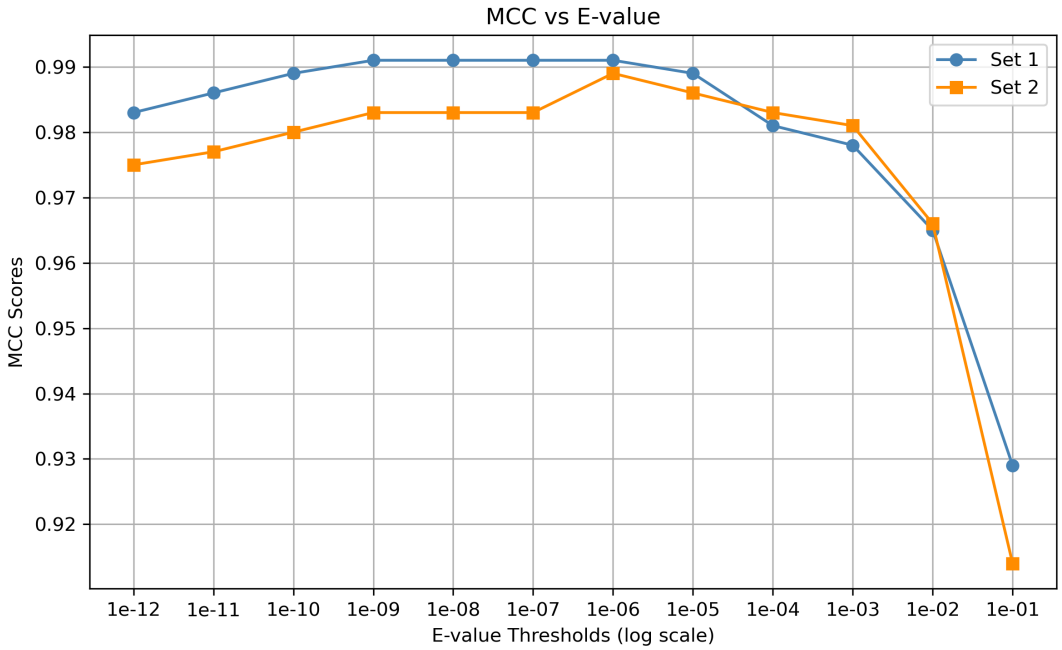


Figure 5: *Graphic representing the trend of MCC 2 over E-value thresholds ranging from* $1e^{-1}$ *to* $1e^{-12}$ *in the two sets displayed in two different colors.*

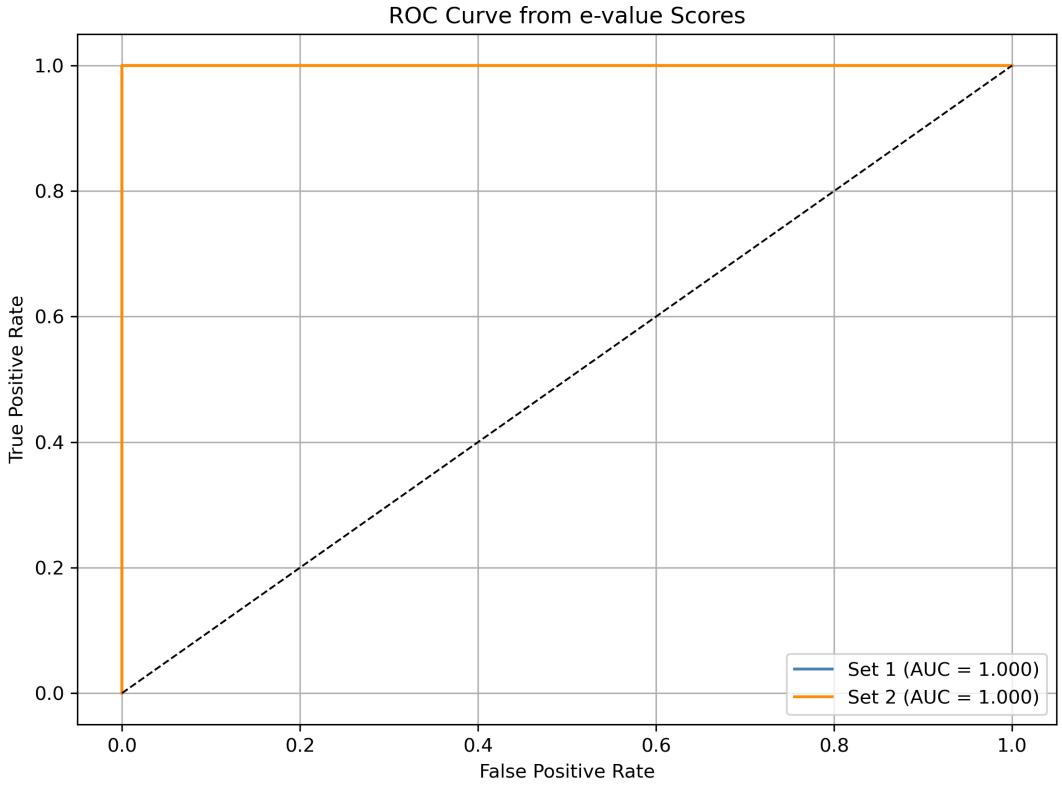


Figure 6: *Graphic of the ROC Curve showing the trade-off between True Positive Rate and False Positive Rate in both sets.*

References

[1] Shiwanthi Ranasinghe and Donald P. McManus. Structure and function of invertebrate kunitz serine protease inhibitors. *Developmental Comparative Immunology*, 39(3):219–227, 2013.

[2] Camila Ramalho Bonturi, Ana Beatriz Silva Teixeira, Vitória Morais Rocha, Penélope Ferreira Valente, Juliana Rodrigues Oliveira, Clovis Macêdo Bezerra Filho, Isabel Fátima Correia Batista, and Maria Luiza Vilela Oliva. Plant kunitz inhibitors and their interaction with proteases: Current and potential pharmacological targets. *International Journal of Molecular Sciences*, 23(9), 2022.

[3] Emma-Karin I. Millers, Manuela Trabi, Paul P. Masci, Martin F. Lavin, John de Jersey, and Luke W. Guddat. Crystal structure of textilinin-1, a kunitz-type serine protease inhibitor from the venom of the australian common brown snake (pseudonaja textilis). *FEBS Journal*, 276(11):2919–2930, May 2009.

[4] Paolo Ascenzi, Alessandro Bocedi, Martino Bolognesi, Andrea Spallarossa, Massimo Coletta, Raimondo De Cristofaro, and Enrico Menegatti. The bovine basic pancreatic trypsin inhibitor (kunitz inhibitor): a milestone protein. *Current Protein & Peptide Science*, 4(3):231–251, June 2003.

[5] Manasi Mishra. Evolutionary aspects of the structural convergence and functional diversification of kunitz-domain inhibitors. *Journal of Molecular Evolution*, 88(7):537–548, 2020.

[6] Sean R Eddy. Hidden markov models. *Current Opinion in Structural Biology*, 6(3):361–365, 1996.

[7] Sean R. Eddy. What is a hidden markov model? *Nature Biotechnology*, 22(10):1315–1316, 2004.