Bioinformatics

Building and testing of a Profile Hidden Markov Model for Kunitz proteins

Marco Centenaro

Department of Pharmacy and Biotechnology (FABIT)

Abstract

Motivation: the aim of this work is the training and performance assessing of a BPTI/Kunitz domain Profile Hidden Markov Model (HMM), starting from a multiple structure alignment of structures found in the PDB, for the annotation of Kunitz proteins.

Results: the Model yielded an MCC = 1, resulting in perfect prediction capabilities with the e-value threshold of 0.002. The model has been tested using a 2-fold cross validation technique and datasets retrieved from the Swissprot section of the Uniprot database. However, there were initial problems with the selection of Kunitz proteins from the Uniprot database, relating to the phenomenon of annotation incompleteness, that, not accounted for in the beginning, resulted in a biased model for which the performances shown in the training and validation lack real-world applicability.

Contact: marco.centenaro@studio.unibo.it

Supplementary material: supplementary material and data are available at **GitHub**

1 Introduction

1.1 Kunitz domain

Peptidases are indispensable for the survival of all kinds of organisms as they break down substrate proteins, but their activities need to be kept under strict control. Inhibitors of peptidases, the protease inhibitors play crucial roles in natural systems by tightly regulating the protease activity and acting as a switch in many signaling pathways (Mishra, 2020).

Kunitz-domain inhibitors known from animal sources are typically of 50–70 amino acids in length and adopt a conserved structural fold with two antiparallel β-sheets and one or two helical regions that are stabilized with three disulfide bridges with the bonding pattern of 1–6, 2–4, 3–5. The disulfide bridges maintain the structural integrity of the inhibitor and also present the protease-binding loop at its surface (Figure 1). A highly exposed P1 active site residue at position 15 is usually arginine or lysine inserts into the S1 site of the cognate protease and is the primary determinant of the specificity of serine protease inhibition. This motif was first identified in the bovine pancreatic trypsin inhibitor (BPTI)-like protease inhibitors, which are strong inhibitors of serine proteases like trypsin and chymotrypsin (Mishra, 2020).

Kunitz proteins are stable as standalone peptides, and work as competitive protease inhibitors in their free form. These properties have led to attempts at developing biopharmaceutical drugs, in particular for the treatment of angioedema with the drugs kallikrein and Ecallantide (Lehmann, 2008) and for the treatment of acute respiratory distress syndrome with the drug Depelestat. The most successful example is Trasylolm a drug (now discontinued) used to reduce bleeding during surgery. Unsuccessful attempts have been made to treat cystic fibrosis (Attucci et al., 2006).

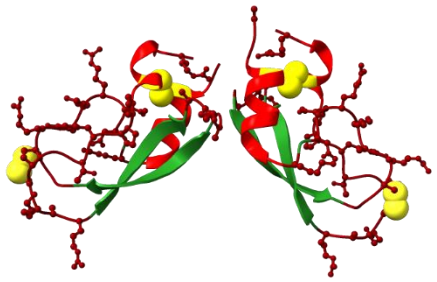

Figure 1: both chains of Bovine Trypsin Inhibitor protein, in yellow are highlighted the signature Kunitz disulfide bonds, helixes in red, strands in green. Image realized with ChimeraX.

1.2 Profile Hidden Markov Model

Hidden Markov Models (HMMs) are statistical trainable models that have been used extensively for modelling protein families with grate success, as they provide high sensitivity. With the advent of deep learning more complex non-Markovian models, HMMs have seen a decline and are no longer the gold standard for protein annotation. However, their robustness, interpretability, ease of train and resource-friendliness, in terms of computational resources and time needed for their development still makes them appears in protein annotation pipelines (Bjerregaard et al., 2025).

One open-source software for modelling HMMs is HMMER3, a program that builds on the original HMMER programs (Eddy, 1998), achieving BLAST comparable speed.

2 Methods

The experiment was conducted using GNU Core Utilities tools for text file manipulation and python3 scripts created specifically for this task. HMM was built using HMMER3 (version 3.3.2). Other third-party tools employed are CD-HIT (version 4.8.1) and the request module for python3. All the commands were executed in the Anaconda environment for easy dependencies management and installation. Plots have been realized with Python. Supplementary information such as the entire list of commands used, version of Anaconda packages used and datasets and files can be found in the GitHub repository at GitHub

2.1 Structure collection and initial filtering

Protein structures were retrieved from the PDB, filtering entries based on the following criteria: resolution ≤ 3.5 Å, PFAM ID PF00014 (Kunitz domain), and sequence length between 45 and 80 residues. The report was downloaded in CSV format and included the following fields:
PDB ID, Entry ID, Polymer Entity ID, Auth Asym ID, sequence data.
A FASTA file was generated from the report using custom awk and grep pipelines. Sequence clustering was performed using CD-HIT, and representative sequences from each cluster were extracted. Sequence length filtering was applied to ensure compliance with the target size range, and one outlier (sequence > 80 residues) was removed.

2.2 Multiple structural alignment

Multiple structure alignment was performed using the PDBeFold web server. PDB IDs and chain identifiers were formatted accordingly, and sequences with outlier RMSD values (e.g., 5JBT with RMSD = 2.9 Å) were excluded to improve alignment quality. The resulting structural alignment was converted into FASTA format using awk and sed, ensuring all residues were in uppercase and headers were cleaned.

2.3 Profile HMM building and dataset construction for validation

The sequence alignment derived from the multiple structural alignment was used to train a profile Hidden Markov Model (HMM) using the HMMER3 package.
To evaluate the model, positive and negative datasets were obtained from UniProt Swiss-Prot:

- Positive dataset: reviewed and annotated Kunitz sequences
- Negative dataset: all remaining reviewed sequences.

Both datasets were cleaned using awk to simplify headers. Cross-referencing with PDB data was performed using a custom Python script to remove any overlap sequences between training and evaluation datasets. Redundant UniProt IDs were filtered using sort and uniq.

The resulting positive and negative identifiers were randomly shuffled using sort -R and split into two halves for 2-fold cross-validation. Sequence extraction was performed using a custom script (get_sequence.py), generating FASTA files for each fold.

2.5 HMMSearch and model evaluation

HMMSearch command was run on the negative and positive datasets with the flags --max and -Z 1000, disabling heuristics and normalizing the results for dataset size comparability. Hits were parsed to generate .class files containing sequence ID, class label (1 for positive, 0 for negative), and e-value. Due to HMMER's default e-value threshold (10), some negative sequences were not included in the output; these were reintroduced in the negative datasets using comm with a placeholder e-value of 10.
Merged datasets were evaluated using a custom Python script (performance.py) that assessed model performance using Matthews Correlation Coefficient (MCC) across varying e-value thresholds. The optimal threshold was selected based on the highest MCC and minimal false positives/negatives.

2.6 Annotation incompleteness correction

Manual inspection revealed that several low-e-value sequences classified as negatives were actually Kunitz proteins annotated by InterPro (IPR036880) but not by PFAM. Additional queries to UniProt confirmed that PFAM annotations did not fully capture the entirety of the Kunitz proteins, placing a bias in the model. Eleven InterPro-only annotated proteins were added to the positive dataset and validation of the model redone.

2.7 Final evaluation and threshold tuning

A final evaluation was conducted using the updated datasets. The model reached MCC = 1 at an optimal threshold of 0.002, with no false positives or false negatives after including missing annotations. The last misclassified protein (Uniprot entry C0HMD5) was identified as a true Kunitz protein (Prosite-annotated) and was reassigned to the positive class.

3 Results

3.1 Initial PDB dataset and PDBeFOLD multiple structural alignment

The initial pool of Kunitz protein structures downloaded from the PDB featured 158 entries, a filtered set of 25 representative sequences was later obtained after running the CD-HIT program with default setting (similarity 0.9 for clustering). One of the 25 sequences (entry 2ODY_F) was removed for being too long (129 residues compared to the 80 resi-

due threshold). After accordingly processing the CD-HIT output as explained in the methods the multiple structural alignment was performed using the PDBeFOLD server. Metrics relative to the alignment are displayed in Figure 2 below.

```
                RESULTS OF MULTIPLE ALIGNMENT

 SUMMARY

                                                    Consensus scores
 ##    Structure                        Nres  Nsse    RMSD  Q-score
  1  PDB 5nx1:C  .....................    54     4   0.9392  0.8433
  2  PDB 6bx8:B  .....................    55     4   0.4957  0.8849
  3  PDB 4bqd:A  .....................    78     6   0.4536  0.6267
  4  PDB 1yc0:I  .....................    66     4   0.5593  0.7321
  5  PDB 4dtg:K  .....................    60     4   0.4290  0.8166
  6  PDB 1f5r:I  .....................    57     4   0.6436  0.8386
  7  PDB 1zr0:B  .....................    63     4   0.5192  0.7706
  8  PDB 3wny:C  .....................    57     4   0.6716  0.8353
  9  PDB 1bun:B  .....................    61     3   0.9960  0.7383
 10  PDB 3m7q:B  .....................    61     4   0.7332  0.7735
 11  PDB 6q61:A  .....................    59     4   0.5388  0.8210
 12  PDB 4ntw:B  .....................    60     4   1.2642  0.7077
 13  PDB 5yv7:A  .....................    60     4   0.6148  0.7997
 14  PDB 1dtx:A  .....................    59     4   0.6083  0.8140
 15  PDB 3byb:A  .....................    58     4   0.4859  0.8400
 16  PDB 4u30:X  .....................    54     3   0.7906  0.8658
 17  PDB 6yhy:A  .....................    59     4   0.5545  0.8195
 18  PDB 1knt:A  .....................    55     4   0.8080  0.8476
 19  PDB 5jb7:A  .....................    56     4   0.5698  0.8618
 20  PDB 5m4v:A  .....................    57     4   0.6412  0.8389
 21  PDB 1yld:B  .....................    56     4   0.6422  0.8537
 22  PDB 1fak:I  .....................    55     4   0.5332  0.8813
 23  PDB 4u32:X  .....................    54     3   0.8675  0.8545

Number of aligned residues 50
Number of aligned SSEs     3
Overall RMSD               1.0063
Overall Q-score            0.5335
```

Figure 2: PDBeFOLD report

3.2 Building of the HMM

The HMM was built starting from the 23 aligned structures. (1 was removed for having a poor RMSD value of 2.9) and has a length of 58 residues, in line with the average length of the Kunitz domain. The HMM itself is available in the GitHub repository. From the sequence logo in figure 3 it clearly appears that the model captured the conservation of the cysteines forming the characteristic disulfide bonds. As well as of other aromatic residues around the middle of the domain.

3.3 Dataset construction for model evaluation

Datasets were initially obtained querying the Uniport Swissprot database for all proteins containing the Pfam annotated Kunitz domain (PF00014) except the ones used for training of the model, obtaining the positive dataset, and the negative dataset was obtained considering all the remaining proteins, however, after running the HMMSearch command and sorting the negative dataset by descending e-value and inspecting the Uniprot entries of the low e-value scoring proteins, 11 of those proteins were revealed to be Kunitz proteins, however they were not annotated by Pfam, but by the Interpro database (Interpro Kunitz Superfamily identifier: IPR036880). Querying the database with the appropriate queries revealed that, considering the Swissprot database, all the Pfam annotated Kunitz proteins are also Interpro annotated, but the contrary is not true. In particular, there are 11 Interpro Kunitz proteins that are not considered as such by the Pfam database, as visualized in figure 4.

To fix this issue, a python script was produced to move those 11 sequences from the negative dataset to the positive one, and model evaluation was conducted using those fixed datasets. Later, another Kunitz protein was found in the same way, this time only annotated by Prosite. This protein was shifted in the positive dataset as well.

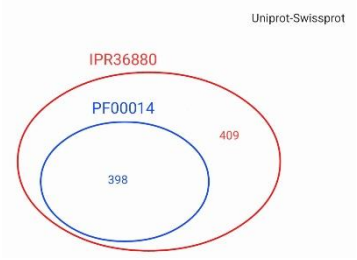

Figure 4: visualization of the annotation state of Kunitz proteins in Swissprot considering Pfam and Interpro cross references

For the two-fold cross validation, the entire negative and positive sequences datasets containing 572820 and 412 entries respectively, were randomized and subsequently halved, following the procedure explained in the methods section.

3.4 Two-Fold cross validation

The HMMSearch command was run on the adjusted datasets in FASTA format to obtain the e-values for each entry, and class files were obtained using an awk script. Performance was assessed on the e-value relative to the whole sequence only since the model is specifically designed to identify whole proteins. Thresholds considered initially ranged from e to $1e^{-13}$ with jumps of 1 order of magnitude (decreasing by 1 the value of the exponent). After finding an initial best threshold of $1e^{-3}$ (approximated to 0.001), it was later refined by multiplying this value with integers ranging from 1 to 10.
The resulting confusion matrices are displayed in figure 5 below.

Table 1: 2 fold cross validation - confusion matrices, threshold 0.001

I Dataset		II Dataset	
True Negatives	False Positives	True Negatives	False Positives
286409	1	286411	0
False Negatives	True Positives	False Negatives	True Positives
1	195	0	197

Table 2: 2 fold cross validation - confusion matrices, threshold 0.002

I Dataset		II Dataset	
True Negatives	False Positives	True Negatives	False Positives
286410	0	286411	0
False Negatives	True Positives	False Negatives	True Positives
0	196	0	197

Figure 5: confusion matrices for the threshold values of 0.001 and 0.002

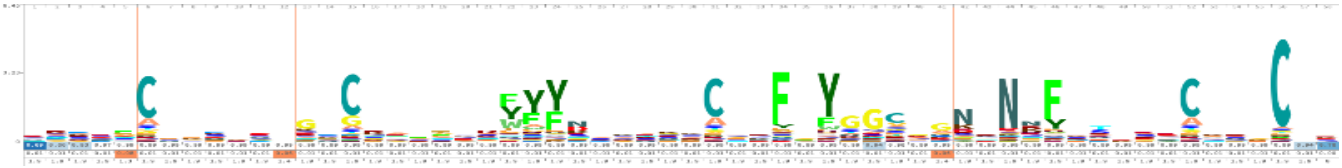

Figure 3: Sequence logo of the Kunitz protein HMM

As a metric for the performance, given the highly unbalanced classes the Matthews Correlation Coefficient (MCC) was used.
With a threshold of 0.002 a MCC of 1 is achieved, suggesting perfect prediction performance of the model.

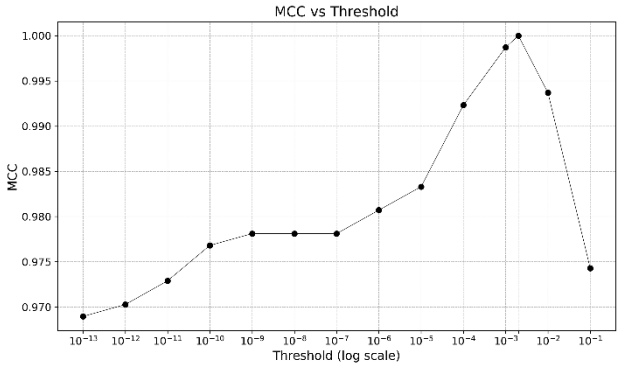

Figure 6: plot showing MCC values (computed on the entire positive and negative datasets) as threshold value increase

4 Discussion

Despite the model's MCC of 1, it is unlikely that it will always yield perfect results, especially considering the problems relating to the precision of the negative and positive datasets due to annotation incompleteness. Considering the high number of domain databases such as Interpro, Pfam and Prosite, and considering that is not always possible to add them in the field of "cross-references" in the Uniprot advanced search (for example, it is not possible to look for proteins with "Prosite ProRule" annotation, exactly the annotation of the C0HMD5 Uniprot entry) it is not possible to say with 100% certainty if there still are or not Kunitz proteins in the negative dataset. Nevertheless, the figure 8 plot shows that there is a gap around the e-value 10^{-3}. The "last" positive protein is the entry C0HMD5, which is an unusual Kunitz protein since it is 163 residue long, with an e-value closer to the first negative proteins. It could be considered an outlier Kunitz, however, the existence of this protein suggests that there may be other similar, unusually long Kunitz, especially considering that the positive dataset was much smaller compared to the negative dataset.
When using this model, proteins with an e-value around 10^{-3} should be manually inspected before deciding if considering them Kunitz protein is recommended.
A ROC curve and AUC score are not available in this work because, due to the very good performance of this model at every threshold, the ROC curve is collapsed on the top left corner of the graph. The plot in figure 6 is provided as a substitute for the ROC curve and is much more readable.

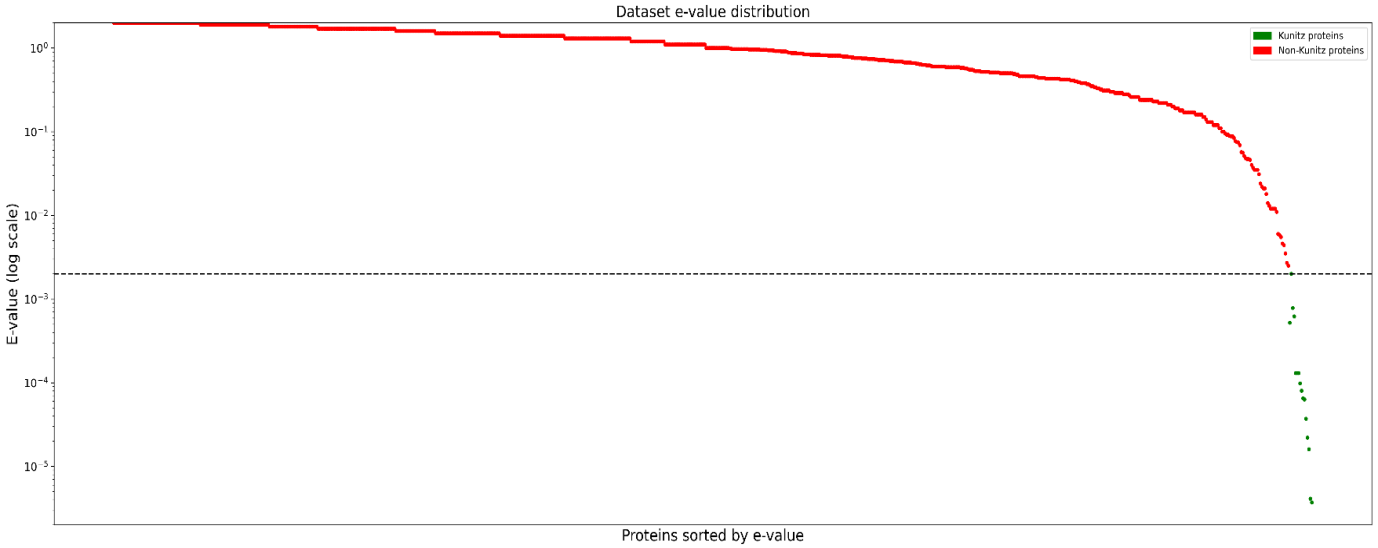

Figure 8: visual depiction of the threshold dividing Kunitz and non-Kunitz proteins in two fields. The segmented line is the threshold value of 0.002

5 Founding

This work was founded by the author himself.
Conflict of interest: no conflict of interest declared.

References

Attucci, S., Gauthier, A., Korkmaz, B., Delépine, P., Martino, M.F.-D., Saudubray, F., Diot, P., Gauthier, F., 2006. EPI-hNE4, a Proteolysis-Resistant Inhibitor of Human Neutrophil Elastase and Potential Anti-Inflammatory Drug for Treating Cystic Fibrosis. J. Pharmacol. Exp. Ther. 318, 803–809. https://doi.org/10.1124/jpet.106.103440

Bjerregaard, A., Groth, P.M., Hauberg, S., Krogh, A., Boomsma, W., 2025. Foundation models of protein sequences: A brief overview. Curr. Opin. Struct. Biol. 91, 103004. https://doi.org/10.1016/j.sbi.2025.103004

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Lehmann, A., 2008. Ecallantide (DX-88), a plasma kallikrein inhibitor for the treatment of hereditary angioedema and the prevention of blood loss in on-pump cardiothoracic surgery. Expert Opin. Biol. Ther. 8, 1187–1199. https://doi.org/10.1517/14712598.8.8.1187

Mishra, M., 2020. Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. J. Mol. Evol. 88, 537–548. https://doi.org/10.1007/s00239-020-09959-9