A HMM-based model for Kunitz domain

Omid Mokhtari*

* Master of Bioinformatics, University of Bologna, Italy

Abstract

Motivation: Kunitz-type protease inhibitors are small domains of proteins involving in diverse functional roles and found in almost all organisms. Hidden Markov models (HMM) provide an efficient way to statistically evaluate protein sequences and classify Kunitz/non-Kunitz domains. The aim of this paper is to develop a model able to detect unannotated members of this protein family.

Results: Testing our model using the whole Swiss-Prot database showed high Matthew correlation coefficient (MCC) score through the performance of binary classification of Kunitz proteins what proves the efficiency of HMM.

Contact: omid.mokhtari@studio.unibo.it

Supplementary information: Supplementary data are available at

1 Introduction

Aprotinin is a single-chain globular polypeptide derived from bovine lung tissue which prevents the breakdown of fibrin protein and helps blood clotting. Hence, used as a drug marketed under the name Trasylol to reduce preoperative blood loss and the need for blood transfusion in patients undergoing cardiac surgery[i].

This protein is a classic member of Kunitz domain, a group of serine protease inhibitors which are one of the most extensively studied protein models. They all share the same structural core; about 50–60 amino acids long, molecular weight of about 6 kDa, folded into a disulfide-rich α/β structure**.** consisting three disulfide bonds: Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51. Lys 15 has a major role in the inhibition of trypsin active site by penetrating its side chain deeply into the binding pocket **(Fig. 1).**[ii,iii]

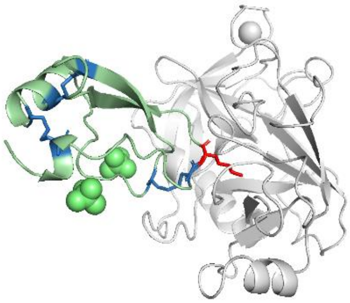
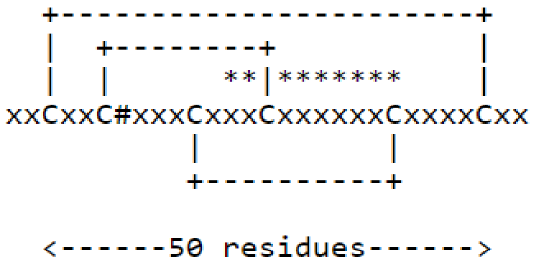


Fig. 1. *Structure of Aprotinin (PDB ID 3TGI). Kunitz chain is shown in green, Lys15 in red and cysteines involved in disulphide bonds in blue.*

Kunitz proteins could contain one or more Kunitz domains, characterized by their conserved pattern of disulphide bonding. Two of these disulphide bridges maintain the native conformation and the third one stabilizes the binding domains **(Fig. 2)**. Kunitz domain inhibits Serine protease with a non-covalent interaction between its protease-binding loop and enzyme's active site, which resembles Michaelis complex of enzyme-substrate without any conformational changes.[iv]

In this work we aim to build a statistical model for discrimination of proteins with Kunitz domain. To do so, we started from members of this protein family that already have available experimental structural information. Then, using multiple structure alignment we derived the corresponding sequence profile based on Hidden Markov Model (HMM) to be the seed of our model. Finally, we evaluated our model and results clearly indicated great performance of HMM.



```
  +-----------------------+
  |  +--------+           |
  |  |      **|*******    |
xxCxxC#xxxCxxxCxxxxxxCxxxxCxx
          |          |
          +----------+

  <------50 residues------>
```

Fig. 2. *Pancreatic trypsin inhibitor (Kunitz) family signature and profile from ProSite*
'C': conserved cysteine involved in a disulfide bond.
'#': active site residue.
'': position of the pattern.*

2 Methods

2.1 Seed Preparation

In order to construct the seed alignment, we extract the highly annotated proteins with available experimental structure. The Pfam and PDB database were used to retrieve the PDB IDs and their associated information on position of the Kunitz domain[v,vi]**.**

Therefor we made a table from available data in Kunitz_BPTI family (PF00014) including UniProt residues, PDB IDs, PDB chains and PDB residues for the Kunitz domain **(Supplementary table 1)** and also performed a PDB search with the following filters; Identifier for Pfam family set to PF00014 to obtain only the proteins containing Kunitz domain, Data collection resolution set to less than 3 and 0 mutation in the polymer entity **(Supplementary table 2).** Since some PDB IDs in Pfam table has multiple Kunitz chains, Unique PDB IDs were kept and then the common chains from both tables were extracted. The length of each domain is expected to be around 50 residues, so we created a plot on distribution of lengths **(Supplementary Fig. 1)** and removed those with less than 40 residues to exclude those not containing the full domain.

Finally, to cure the retrieved Kunitz Domain, we downloaded the sequence and PDB file associated to each chain and we trimmed them based on the position of residues of domains in the Pfam table using python scripts. In order to remove the redundancy and avoid bias in our model we performed clustering using CD-Hit[vii] with the threshold of 95% and recommended word size of 5. Since we trimmed all the sequences and restrict. The representative of each cluster was chosen as the longest one because we already trimmed them based on the Kunitz domain.

2.2 Model generation

The structures of the selected sequences were used to perform a multiple structural alignment and derive the corresponding sequence alignment for building the model. For this purpose, we utilized mTM-align[viii]. This algorithm builds a multiple structural alignment progressively based on the pairwise structure alignment generated by highly efficient TM-align. The resulted sequence alignment was further used to build our profile HMM model using hmmbuild function in HMMER v3.3 (hmmer.org).

2.3 Model evaluation

In order to evaluate our model, we created two datasets:

1. A negative set composed by all non-Kunitz entries of UniProtKB/Swiss-Prot containing 566628 sequences
2. A positive set composed by all entries of UniProtKB/Swiss-Prot that contains Kunitz domain. Since we used some of entries in this dataset for model construction, we built a query excluding those that have PDB structure associated to them resulting in 336 sequences.

We labeled the datasets according to the positive and negative sets and shuffled them both. Then we split each set into two portions to use cross-validation resampling method. So, we can train and assess the model with different E-value thresholds on different iterations.

To analyze the sequences, hmmsearch command was used as provided in the bash file in supplementary materials with the following options; To normalize the output e-value, a constant number of 1 is set to avoid different per-sequence e-value calculation. However, we turned off all filters to retrieve the maximum number of outputs but more than half of sample in the negative set were missing so we added them manually to our evaluated tables with a random high e-value of 10. Then we merged each negative and positive portion to make two separate sets and the e-value threshold of each set was optimized to retrieve the maximum accuracy (ACC) and Matthew correlation coefficient (MCC). The comprehensive final threshold was computed as the average threshold of each set.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

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

Formula of Accuracy and Matthew correlation coefficient. where TP = True positive; FP = False positive; TN = True negative; FN = False negative

3 Results

3.1 Seed sequences

Primary number of sequences retrieved from Pfam and PDB query were 321 and 135 respectively. However, after curation and filtering procedure, only 31 entries were used for the seed construction. **(Supplementary table 3)**

The mTM-align algorithm yielded a good alignment of the core as α/β folds were almost perfectly conserved and flexibility of the loop regions could be observed. **(fig. 3)** The inferred sequence alignment also indicted good conservation among cysteine bridges and 15th residue. **(fig. 4)**

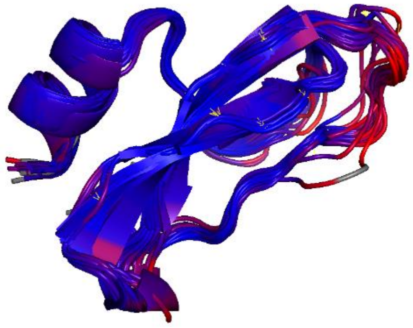


Fig. 3. Superimposition of all 31 structures colored by their RMSD score with respect to PDB Id 5YHY. Image obtained by PyMol (Schrödinger, L. & DeLano, W., 2020.)

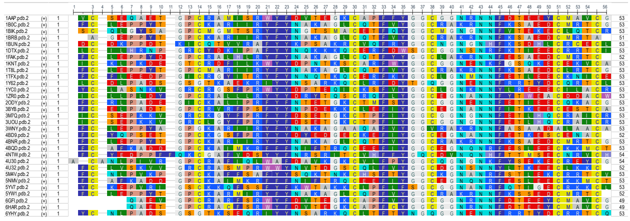


Fig. 4. Multiple sequence alignment of the seeds. image obtained from NCBI Multiple Sequence Alignment Viewer v1.22

3.2 HMM Model

Given the profile of multiple sequence alignment, hmmbuild was able to construct a model capable of describing the position-based emission and transmission probability. Model's logo is shown in figure 5. This algorithm is also able to overcome the problem of low complexity regions by generating a model based on average match state emission probability.

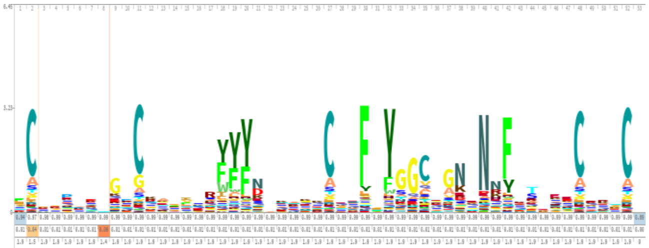


Fig. 5. HMM logo of our model generated by skylign. (Wheeler et al., 2014)

3.3 Model validation

By creating a reviewed negative and positive sets containing 566629 and 336 entries and splitting them to 2 folds we assessed our model based on accuracy and Matthew correlation coefficient. On each iteration we obtained the optimal threshold of 1.8e-9.

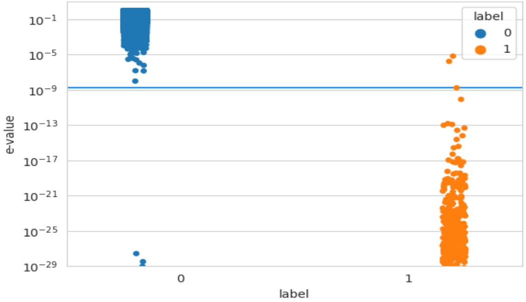


Fig. 6. Distribution of e-values with respect to their negative/positive label. The overall threshold is shown in blue line

The overall validation of our model showed accuracy of 0.999 and MCC equal to 0.991 with 2 false negatives and 4 false positives.

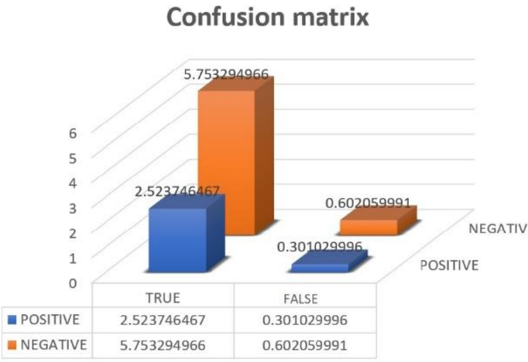
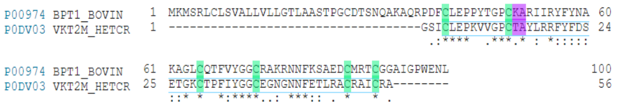


Fig. 7. Confusion matrix in logarithmic scale

Discussion

Despite the very good performance of our model few errors occurred as shown in the confusion matrix **(fig. 7)**. P0DV03, P0DV04, P0DV05, P0DV06 were 4 false positives with e-value range of 1e-28-1e-29. These proteins belong to *Heteractis crispa* organism and all of them annotated with same PROSITE family signature; PS00280 and PS50279; figure 8 demonstrates the conservation among the cysteine residues and the important sites with one the typical examples of Kunitz domain. Thus, we can be sure that the labels are wrong and proteins lack Pfam annotation. D3GGZ8 and O62247 were the only false positives with the e-value of 1e-06 which may suggest poor training data of our seed or wrong Pfam annotation. The hmmsearch alignment couldn't explain the conservation of Kunitz domain, moreover, we downloaded Alphafold structure of O62247 and found all polar interactions between cysteine residues but no meaningful disulphide bound was found **(supplementary fig. 2)**.



```
P00974 BPT1_BOVIN    1 MKMSRLCLSVALLVLLGTLAASTPGCDTSNQAKAQRPDFCLEPPYTGPCKARIIRYFYNA  60
P0DV03 VKT2M_HETCR    1 ------------------------------------GSICLEPKVVGPCTAYLRRFYFDS  24
                                                            .:****  .***.* : *:::::

P00974 BPT1_BOVIN   61 KAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGAIGPWENL                    100
P0DV03 VKT2M_HETCR  25 ETGKCTPFIYGGCEGNGNNFETLRACRAICRA--------                     56
                       ::* *  *:****..: ***:: . *   * .
```

Fig. 8. Alignment of P00974 and P0DV06 obtained from UniProt.

[i] Judith L Kristeller, Brian P Roslund, and Russell F Stahl, "Benefits and Risks of Aprotinin Use during Cardiac Surgery," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 28, no. 1 (2008): 112–24.

[ii] Izabela Krokoszynska, Michal Dadlez, and Jacek Otlewski, "Structure of Single-Disulfide Variants of Bovine Pancreatic Trypsin Inhibitor (BPTI) as Probed by Their Binding to Bovine β-Trypsin," *Journal of Molecular Biology* 275, no. 3 (1998): 503–13.

[iii] Ruiming Zhao et al., "SdPI, the First Functionally Characterized Kunitz-Type Trypsin Inhibitor from Scorpion Venom," *PloS One* 6, no. 11 (2011): e27548.

[iv] Shiwanthi Ranasinghe and Donald P McManus, "Structure and Function of Invertebrate Kunitz Serine Protease Inhibitors," *Developmental & Comparative Immunology* 39, no. 3 (2013): 219–27.

[v] Helen M Berman et al., "The Protein Data Bank," *Nucleic Acids Research* 28, no. 1 (2000): 235–42.

[vi] Jaina Mistry et al., "Pfam: The Protein Families Database in 2021," *Nucleic Acids Research* 49, no. D1 (2021): D412–19.

[vii] Limin Fu et al., "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data," *Bioinformatics* 28, no. 23 (2012): 3150–52.

[viii] Runze Dong et al., "MTM-Align: An Algorithm for Fast and Accurate Multiple Protein Structure Alignment," *Bioinformatics* 34, no. 10 (2018): 1719–25.