Laboratory of Bioinformatics 1 Module B

Annotation of the Kunitz domain using Hidden Markov Models

Rosaria Tornisiello[1,*]

[1]International Master in Bioinformatic, University of Bologna.

*To whom correspondence should be addressed.

Abstract

Motivation: The Kunitz domain, also known as bovine pancreatic trypsin inhibitor (BPTI) is one of the most extensively studied globular proteins since its use during surgical interventions reduces hemorrhagic complications[1]. Moreover, it turned out to be a useful molecule for studying molecular recognition and protein/protein interactions[1]. The aim of this work is to build and test a Hidden Markov Model (HMM) for the classification of proteins containing the Kunitz domain. This method could allow the recognition of new BPTI-containing proteins.

Results: The HMM is found to be an excellent predictor of Kunitz-type proteins, yielding a high accuracy and a good Matthews Correlation Coefficient (MCC). The prediction is wrong solely in few interesting cases that point out some uncertainty in the SwissProt annotation.

Contact: rosaria.tornisiello@studio.unibo.it

Supplementary information: Supplementary data are available at GitHub - RosariaTornisiello/KunitzProject_TornisielloRosaria.

1 Introduction

Some protease inhibitors are characterized by a specific domain which is referred to as Kunitz domain. Examples of Kunitz-type proteins are: aprotinin (BPTI), Alzheimer's Amyloid Precursor protein (APP), and Tissue Factor Pathway Inhibitor (TFPI)[2]. BPTI has been one of the most extensively studied globular proteins in the early years of structural biology since it is a useful tool for analysing protein conformation, folding and dynamics[1]. Since it has a high specificity for plasmin it is used as antihemorrhagic drug during cardiopulmonary surgery and orthotopic liver transplantation[1].

The BPTI primary structure is composed by 58 amino acid residues and is stabilized by three disulfide bridges with the bonding pattern C1-C6, C2-C4 and C3-C5[3]. It shows high conservation of the Lys/Arg15 and Cys residues (at positions 5, 14, 30, 38, 51, and 55) which are involved in the stabilizing disulfide bridges[1].

Different crystal forms of monomeric BPTI have been isolated and studied by both X-ray crystallography and Nuclear Magnetic Resonance (NMR). Its structure is very compact and pear-shaped[1]. The conserved site Lys/Arg15 contains the reactive bond of this domain and is located at the narrow end of the molecule[1]. BPTI domain contains a hydrophobic core composed by the side chains of the following residues: Phe4, Cys5-Cys55, Phe22, Tyr23, Cys30-Cys51, Phe33, Tyr35, and Phe45[1]. Three hydrogen bonds from the side chain of Asn43 to main chain atoms play an important role in stabilizing the molecule[1]. The BPTI secondary structure is composed by two helical regions: one and one-half turns of 3_{10} helix near the N-terminus and almost three turns of α-helix are found near the C-terminus[1]. Finally, it contains a long-twisted double-stranded antiparallel β-loop from Ile18 to Tyr35 and an antiparallel β-sheet hydrogen bonding is also made between residues Phe45 and Tyr21[1]. The segment responsible for protease inhibition is called the protease-binding loop[3]. This loop binds tightly to the concave active site of the enzyme given its high complementary[3].

Given the important role of the Kunitz-type proteins, being able to recognize them could be advantageous. This work describes the generation of a Hidden Markov Model (HMM) and its implementation for the purpose of the binary classification of proteins archived in SwissProt[4], in containing and non-containing the Kunitz domain.

2 Methods

2.1 Generation of the HMM

The first step is the selection of protein structures annotated as Kunitz-type (PF00014) by the Pfam[5] database and with high resolution (<3 Å) by means of the advanced search of RCSB PDB[6]. The retrieved structures are

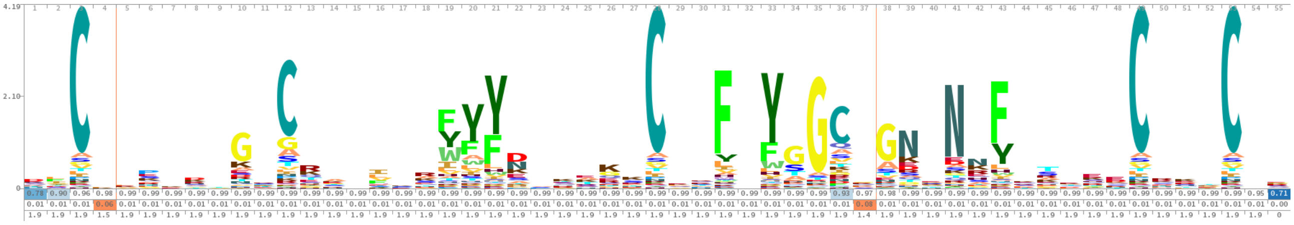


Fig. 3 HMM logo obtained using Skylign.

aligned with a manually selected Kunitz-type structure (3TGI chain I) in pairwise modality in order to select solely the chains containing the Kunitz domain. This step is performed using PDBeFold[7]. To obtain a non-redundant set, the selected sequences are clustered using CD-HIT[8] that uses a threshold of percentage of sequence identity equal to 90% by default. The next step is the manual selection of a representative structure for each cluster and the removal of the mutants. The chosen proteins are aligned in a multiple structural alignment by means of PDBeFold. The output is screened to evaluate the root mean square deviation (RMSD) of each protein and the structural alignment is visualized using PyMol[9]. The multiple sequence alignment is performed using Jalview[10]. Proteins with RMSD higher than 1 that seemed not to align in a proper way with the others are removed. The resulting sequence alignment is used to build an HMM by the use of the *hmmbuild* tool of HMMER[11] (version 3.3). The HMM logo is obtained using Skylign[12].

2.2 Dataset production

The positive dataset, containing only reviewed proteins annotated by Pfam as Kunitz-type, and a negative dataset composed by reviewed proteins non Kunitz-type, are retrieved through the advanced search of UniProt[4]. In order to exclude from the positive set the proteins contained in the seed alignment used to generate the HMM profile, their related UniProt ID is retrieved using the Retrieve/ID mapping tool[13]. Then the IDs are inserted in the query to exclude them from the positive dataset. In order to remove further identical sequences from the positive dataset, BLAST[14] is used to retrieve proteins 100% identical to the ones included in the seed alignment, finally the latter are removed from the positive dataset using a python script. The positive and negative datasets are merged in a unique file.

2.3 Screening of the dataset

The final dataset is screened with the HMM profile previously generated via the *hmmsearch* tool of HMMER, setting the Z parameter equal to 1 to normalize the calculation of the e-value considering the difference in the number of sequences composing the positive and negative sets. Subsequently, the hmmsearch output is used to obtain two pandas[15] dataframes. Each dataframe is composed of two columns for each UniProt ID containing the E-value (full sequence in one case and best domain in the other one) and 1 or 0 depending on the presence or absence of the Kunitz domain, respectively. Finally, since the hmmsearch tool implements an inclusion threshold of 10, not all the entries of the dataset are included in its output, consequently, their E-value is manually set to 10.

2.4 2-fold cross validation

Both the datasets are handled using a python script developed in house (available online). Each dataset is split in a train and a test subset. For both the train sets (full sequence and best domain E-value) the threshold is set in a range from 10^{-20} to 1. Proteins are considered predicted as positives (Kunitz-containing) if their E-value is lower than the threshold, otherwise they are considered predicted as negatives. Moreover, the Matthews Correlation Coefficient (MCC) and the accuracy (ACC) is computed for each threshold, according to the formulas shown below. The best E-value threshold is chosen for both the train datasets, based on the higher MCC and accuracy. The two thresholds are then applied on their respective test datasets. Finally, the accuracy, the confusion matrix, the Receiving Operating Characteristic curve (ROC curve) and the Area Under the Curve (AUC) are computed. The whole workflow is repeated switching the train and the test datasets in a 2-fold cross validation.

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$
$$MCC = TP * TN - FP * FN/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

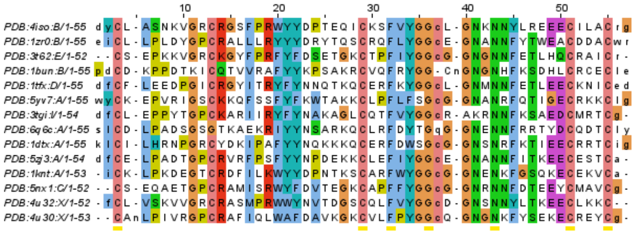
3 Results and discussion

3.1 Seed alignment and HMM profile generation

The advanced search in RCSB PDB gives as output 235 entries. The pairwise alignment of the selected proteins with 3TGI:I using PDBeFold returns 258 chains out of 235 structures. The resulting chains are downloaded in fasta format and the file is given as input to CD-HIT which clusters them into 20 groups. For each cluster one representative is chosen based on the classification made by CD-HIT. The selected structures are manually screened, removing 5JB7 which is a variant, 1FAK, 6BX8 and 1YLD that are mutants. After aligning the resulting structures using PDBeFold, checking the RMSD, further two chains are removed: 4NTX:B and 1D0D:A which show an RMSD higher than 1 and they do not align in a proper way with the other structures. The remainder 14 structures are: 4ISO:B,1ZR0:B, 3T62:E, 1BUN:B, 1TFX:D, 5YV7:A, 3TGI:I, 6Q6C:A, 1DTX:A, 5ZJ3:A, 1KNT:A, 5NX1:C, 4U32:X, 4U30:X. The multiple sequence alignment is shown in figure 1 and the multiple structural alignment is shown in figure 2. The HMM profile generated from this seed alignment is represented as logo in figure 3. The logo shows the conservation of important amino acid residues such as the Cysteines involved in the stabilizing disulphide bonds. The output of PDBeFold is shown in table S1 in the supplementary materials.

3.2 Dataset generation

Employing the advanced search of UniProt, both the positive and the negative datasets are generated. The first one is composed by 339 sequences (after the cleaning step) and the second by 561894 sequences. The two datasets are merged and screened by the HMM profile. The hmmsearch output is used to generate two pandas dataframe characterized by 562233 rows and 2 columns. Each dataframe is then split in two sections, one for the training and one for the testing of the model.



```
PDB:4iso:B/1-55   dyCL-ASNKVGRCRGSFPRWYYDPTEQICKSFVYGGcL-GNKNNYLREEECILACrg
PDB:1zr0:B/1-55   eiCL-LPLDYGPCRALLLRYYYDRYTQSCRQFLYGGcE-GNANNFYTWEACDDACwr
PDB:3t62:E/1-52   --CS-EPKKVGRCKGYFPRFYFDSETGKCTPFIYGGcG-GNGNNFETLHQCRAICr-
PDB:1bun:B/1-55   pdCD-KPPDTKICQTVVRAFYYKPSAKRCVQFRYGG-CnGNGNHFKSDHLCRCECle
PDB:1tfx:D/1-55   dfCF-LEEDPGICRGYITRYFYNNQTKQCERFKYGGcL-GNMNNFETLEECKNICed
PDB:5yv7:A/1-55   wyCK-EPVRIGSCKKQFSSFYFKWTAKKCLPFLFSGcG-GNANRFQTIGECRKKClg
PDB:3tgi:I/1-54   dfCL-EPPYTGPCKARIIRYFYNAKAGLCQTFVYGGcR-AKRNNFKSAEDCMRTCg-
PDB:6q6c:A/1-55   sICD-LPADSGSGTKAEKRIYYNSARKQCLRFDYTGqG-GNENNFRRTYDCQDTCly
PDB:1dtx:A/1-55   kICI-LHRNPGRCYDKIPAFYYNQKKKQCERFDWSGcG-GNSNRFKTIEECRRTCig
PDB:5zj3:A/1-54   dfCE-LPADTGPCRVRFPSFYYNPDEKKCLEFIYGGcE-GNANNFITKEECESTCa-
PDB:1knt:A/1-53   -iCK-LPKDEGTCRDFILKWYYDPNTKSCARFWYGGcG-GNENKFGSQKECEKVCa-
PDB:5nx1:C/1-52   --CS-EQAETGPCRAMISRWYFDVTEGKCAPFFYGGcG-GNRNNFDTEEYCMAVCg-
PDB:4u32:X/1-52   -fCL-VSKVVGRCRASMPRWWYNVTDGSCQLFVYGGcD-GNSNNYLTKEECLKKC--
PDB:4u30:X/1-53   --CAnLPIVRGPCRAFIQLWAFDAVKGKCVLFPYGGcQ-GNGNKFYSEKECREYCg-
```

Fig. 2 Multiple sequence alignment obtained by means of Jalview.

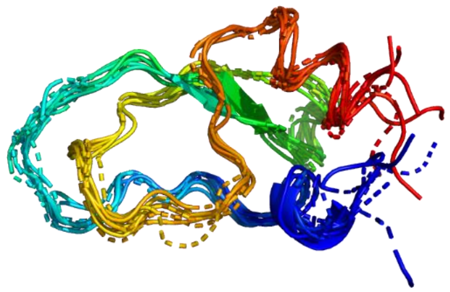


Fig. 2 Multiple structural alignment obtained by means of PyMol.

3.3 Cross validation results

The model generated following this workflow is characterized by an accuracy almost equal to 100 % considering both a full-sequence E-value and a best domain E-value threshold of $1 \cdot 10^{-9}$. Switching the test and the train datasets, the threshold that maximizes the accuracy and the MCC remains the same. The global confusion matrixes are reported as sum in table 1 and table 2. The thresholds, MCC, accuracy and area under the ROC curve are reported in table 3. The scatterplots and the ROC curves are available in the supplementary materials in figure S1. The MCC at different threshold are reported in table S2. The results of full sequence E-value and best domain E-value are basically identical since most of the proteins listed in the positive set are monodomain proteins, so they contain only the Kunitz domain. The selected thresholds applied both to the full sequence E-value and best domain E-value test sets yield the same false negatives and false positives:

- False negatives: D3GGZ8 and O62247, that correspond to the same protein, called BLISTER-5 that is annotated by Pfam as Kunitz-type (PF00014) but not recognized by the HMM profile. D3GGZ8 is expressed by *Haemonchus contortus* and is poorly annotated (inferred from homology) instead O62247 is expressed by *Caenorhabditis elegans* and has a high annotation score (experimental evidence at protein level) . This two proteins are orthologues that appears to have serine protease activity in vitro but is uncertain if this activity is genuine since BLISTER-5 lacks all the catalytic features of serine proteases[16]. Therefore, it is conceivable that the protein sequence underwent changes that caused loss of function. This may be one of the possible reasons why the HMM profile does not succeed in their classification.
- False positives: P56409 called Ornithodorin and P84555 annotated as Kunitz-type by InterPro[18] (IPRO36880). G3LH89 called Bi-KIT, is a strong outlier and is annotated as Kunitz-type both by InterPro and by PROSITE. None of these three proteins is classified by Pfam as Kunitz-type even if they contain the domain, consequently, they are recognized as false positives.

PREDICTED VALUES	ACTUAL VALUE	
	0	1
0	561892	2
1	2	337

Tab. 1 Global confusion matrix for best domain E-value, computed as sum.

PREDICTED VALUES	ACTUAL VALUE	
	0	1
0	561877	5
1	1	354

Tab. 2 Global confusion matrix for full sequence E-value, computed as sum.

	Best domain e-value	Full sequence e-value
Th	$1\cdot10^{-9}$	$1\cdot10^{-9}$
MCC	0.9956	0.9956
ACC	0.9999	0.9999
AUC	0.9999	0.9999

Tab. 3 Threshold (Th), Matthews Correlation Coefficient (MCC), accuracy (ACC), area under the ROC curve (AUC) reported as means of the values obtained from the 2-fold cross validation.

5 Conclusions

This work demonstrates that HMM profiles perform well in the binary classification of domains if they are obtained from a high quality multiple structural alignment. The only cases in which this approach fails are due to inconsistencies in the databases annotation of the specific domain under examination.

6 References

1. Ascenzi, P. *et al.* The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein. *Curr. Protein Pept. Sci.* **4**, 231–251 (2005).

2. Kunitz domain - Wikipedia. https://en.wikipedia.org/wiki/Kunitz_domain.

3. Ranasinghe, S. & McManus, D. P. Structure and function of invertebrate Kunitz serine protease inhibitors. *Dev. Comp. Immunol.* **39**, 219–227 (2013).

4. UniProt. https://www.uniprot.org/.

5. Pfam: Home page. http://pfam.xfam.org/.

6. RCSB PDB: Homepage. https://www.rcsb.org/.

7. PDBe < Fold < EMBL-EBI. https://www.ebi.ac.uk/msd-srv/ssm/.

8. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

9. Schrödinger, LLC. *The {PyMOL} Molecular Graphics System, Version~1.8.* (2015).

10. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Sequence analysis Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinforma. Appl. NOTE* **25**, 1189–1191 (2009).

11. HMMER. http://hmmer.org/.

12. Skylign. https://skylign.org/.

13. Retrieve/ID mapping. https://www.uniprot.org/uploadlists/.

14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

15. pandas - Python Data Analysis Library. https://pandas.pydata.org/.

16. 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.* **169**, 1–11 (2010).

17. ExPASy - PROSITE. https://prosite.expasy.org/.

18. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, (2018).

Supplementary materials

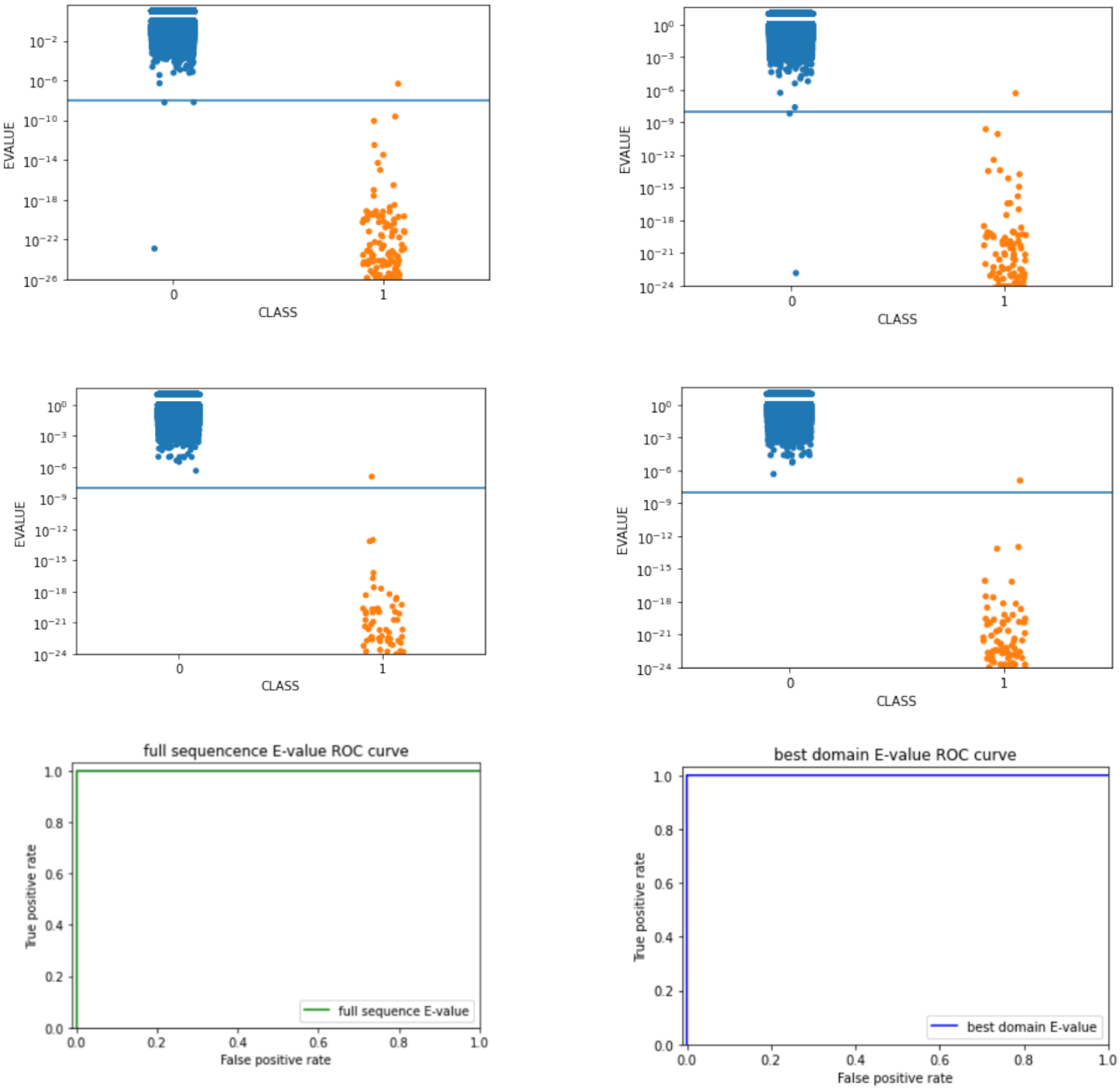


Fig. S1 This figure shows scatterplots obtained in the 2-fold cross validation and the ROC curve for the full sequence e-value on the left-handed side and the ones for best domain e-value on the right-handed side.

PDB ID	NRES	RMSD
4ISO:B	60	0.4948
1ZR0:B	63	0.4561
3T62:E	54	0.7110
1BUN:B	61	1.1993
1TFX:D	58	0.5939
5YV7:A	60	0.6501
3TGI:I	56	0.7808
6Q6C:A	58	0.5179
1DTX:A	59	0.6505
5XJ3:A	56	0.5413
1KNT:A	55	0.7223
5NX1:C	54	0.8137
4U32:X	54	0.7241
4U30:X	54	0.7480

Table S1 This table lists the proteins selected for generating the seed alignment. For each protein in the first column are listed the PDB IDs, in the second column the number of residues (NRES) and in the third the RMSD.

	Full sequence e-value	
	Th	**MCC**
1st-fold	10e-8	0.9968
	10e-9	0.9968
	10e-10	0.9968
	10e-11	0.9968
2nd-fold	10e-8	0.9889
	10e-9	0.9944
	10e-10	0.9916
	10e-11	0.9888

	Best domain e-value	
	Th	**MCC**
1st-fold	10e-8	0.9968
	10e-9	0.9968
	10e-10	0.9968
	10e-11	0.9968
2nd-fold	10e-8	0.9916
	10e-9	0.9944
	10e-10	0.9916
	10e-11	0.9888

Table S2. These tables show the MCC at different thresholds both for full sequence e-value and best domain e-value.