A profile hidden Markov model for predicting Kunitz-type protease inhibitor domains generated by HMMER.

Edoardo Bettazzi

E-mail: edoardo.bettazzi@studio.unibo.it

Supplementary material: https://github.com/EdoardoBettazzi/HMM-for-Kunitz

Abstract

It is described here an approach for building a profile hidden Markov model (profile HMM) for the detection of the Kunitz/BPTI (Bovine pancreatic trypsin inhibitor) domain. The model is generated by the standard HMMER package and trained on structural data. It is optimized via a 2-fold cross-validation procedure, finally detecting the target domain with an MCC (Matthew Correlation Coefficient) of 0.990457, and an accuracy of 0.999976.

---

1. Introduction

Kunitz type proteins are an important group of ubiquitous protease inhibitors, found spanning the evolutionary tree from microbes to mammals. They can have single or multiple Kunitz inhibitory domains, linked together or associated with other domain types [1].

Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI).

Bovine pancreatic trypsin inhibitor is the classic member of this family of proteins and was the first Kunitz-type protease inhibitor described [2].

The canonical motif has a peptide chain of around 60 residues, and it is characterized by a conserved spacing between their cysteine residues, and a typical disulphide bonding pattern [3].

In detail, the overall domain presents a disulphide-rich α + β fold that is stabilized by three highly conserved disulphide bridges with the bonding patterns C1–C6, C2–C4, and C3–C5 (Fig. 1). Two of the disulphide bonds (C1–C6 and C3–C5) are required for the maintenance of native conformation [4] whereas the third (C2–C4) stabilizes the two binding domains [3].

Functionally, Kunitz-domain inhibitors are known to be involved in various physiological processes such as host defense against microbial infection, blood coagulation, fibrinolysis, and inflammation by exhibiting inhibition of serine proteases (e.g trypsin/chymotrypsin/elastase/kallikrein) [1].

Kunitz-domain inhibitors are also known as frequent components of venoms from poisonous animals acting as ion channel blockers [5]. As such, Kunitz-type toxins (KTTs) have developed the ability to block ion channels besides their original function of serine protease inhibition.

While the physiological role of these inhibitors is to resist prey proteases from degrading their venom protein toxins, they have also a recognized therapeutic potential, as the anti-tumoral effect they can exert: the promising pharmacological potential of such inhibitors has been demonstrated in murine renal cell carcinoma model [6] and human glioblastoma cells [7]. Moreover, the Kunitz domain has acquired proximate attention in protein engineering efforts, in order to create specific protease inhibitors relevant for therapeutic applications [8].

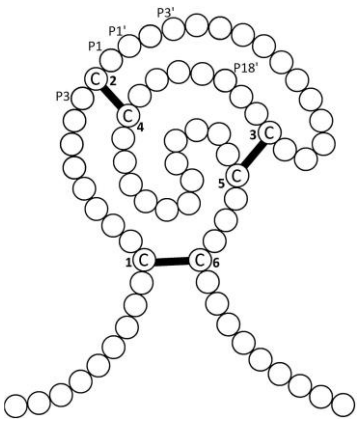

Fig. 1 | Predicted folding scheme for a single domain Kunitz inhibitor showing the characteristic six cysteine residues, three disulphide bonds and protease binding loop (from Ranasinghe and McManus, 2013).

This work aims to provide a basic workflow for the construction of a profile hidden Markov Model for the detection of the Kunitz/BPTI domain in protein sequences. The process starts with the curation of a

training dataset, built from structural data, and a validation dataset, built from sequencing data. The training dataset is used to construct the statistical model, which is then optimized via a 2-fold cross-validation procedure on the validation dataset.

Finally, the optimized model performance is critically investigated to explain any missing ground-truth hit.

2. Methods

2.1 Dataset curation

Training dataset

For the construction of the training dataset, the PDB database [9] was queried to retrieve the IDs of all the proteins, satisfying the following criteria:

- Kunitz domain annotation: according to PFAM code PF00014.
- Resolution method: X-RAY diffraction.
- PDB structure resolution: =< 3 angstrom.
- Absence of mutations in the polymer entity.

As a second source, from the PDBeFold [10] another set of protein sequences, aligning with a given seed structure (3TGI, chain I), was retrieved by pairwise structural alignment against the entire PDB database. The summary was filtered for the quality of structural models, with the following criteria:

- Z-score > 3.
- RMSD < 1.5 angstrom.

The sets from PDB and PDBeFold were merged, and the sequences related to the selected structures were retrieved. To avoid redundancy, a clusterization procedure was performed on the merged set of sequences via CD-HIT v4.8.1 [11], with a 0.95% identity threshold.

Finally, a multiple structural alignment was performed via PDBeFold on the clustered set of proteins. The resulting aligned sequences composed a training set of 21 sequences, used to construct the hmm model.

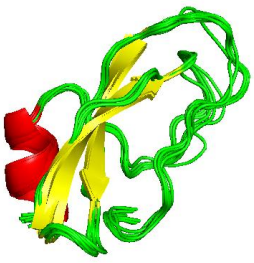

Fig. 2. Superposition of the seed structures at the core level. Image obtained with PyMol [12].

Validation dataset

For the construction of the validation dataset both, a ground-truth positive set of sequences and a ground-truth negative set of sequences, were collected.

For the positive dataset, the UniprotKB database (release 2022_01) [13] was queried to retrieve all the protein sequences, satisfying the following criteria:

- Kunitz domain annotation: according to PFAM code PF00014.
- Reviewed entries.
- Not included in the PDB (to avoid overlapping with the training set).

While for the negative set, the criteria were the following:

- Kunitz domain annotation: according to PFAM code PF00014.
- Reviewed entries.
- Sequence length: 40 to 10000 res.

The sets, totally counting 336 positives and 557,267 negatives, were randomly shuffled and split for applying the 2-fold cross-validation procedure, in order to optimize the model.

2.2 HMM construction

The model was constructed, and trained on the training dataset, with the hmmbuild function from the HMMER tool [14].

2.3 Validation and optimization

The model was tested running the hmmsearch function from HMMER tool on the validation dataset, previously split for applying the cross-validation procedure.

Apart from those modifying the output format, the function was set to operate with the following parameters:

- --max: to turn off any heuristic that would cut off the analysis distantly related proteins.
- -Z: for normalizing the e-value in output.
- --domZ: for normalizing the domain e-value in output.

The performance metrics were computed in relation to the predicted e-value and ground-truth classification of each protein, for a set of e-value threshold values across a range of logarithmic intervals.

In the scope of optimization, the performance metrics were evaluated on both sets, and compared to find the threshold value giving the most optimal metrics.

The chosen metrics of reference for the evaluation were the MCC (Matthew Correlation Coefficient, eq. 1) and the accuracy (eq. 2), defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where:

TP is the number of true positives

TN is the number of true negatives

FP is the number of false positives

FN is the number of false negatives

Comparing the prediction and the ground-truth classification, a set of false results was retrieved and manually investigated.

Please refer to the supplementary material for more details on preprocessing and processing procedures.

3. Results and discussion

The produced model detects the target domains with an acceptable performance. From cross-validation, the most optimal e-value threshold results as 1e-08, giving an accuracy score of 0.99998 and an MMC of 0.99046. The associated confusion matrix is reported in table 1. The HMM logo representation can be found below (Fig. 3).

Table 1. Confusion matrix

	True positives	True negatives
Predicted positives	334	4
Predicted negatives	2	258044

The model reports 6 apparently false results, divided as 4 false positives (UniprotKB IDs: P0DV05, P0DV03, P0DV04, P0DV06), and 2 false negatives (O62247, D3GGZ8). Each result was manually investigated to trace back the cause of false classification.

O62247, D3GGZ8 are predicted as negatives, but not annotated as such: the ground-truth classification of D3GGZ8 is dependent on the other false negative entry, O62247, used as a seed for its annotation and thus considered here a proxy for understanding the misclassification.

O62247 is annotated as positive since it appears to have serine protease activity in vitro, although it is uncertain if this activity is genuine, as the protein lacks all the catalytic features of serine proteases [15]. It can be considered a borderline case, until further investigation will shed new lights on the relation between its structure and its function.

P0DV05, P0DV03, P0DV04, P0DV06 are all isoforms of the same protein, whose sequence actually contains a Kunitz domain, as per PROSITE annotation (PRU00031), and thus it is rightfully predicted as positive. Those 4 sequences are reported as false results because the relative UniprotKB entry does not present the PFAM identifier PF00014, used for the construction of the datasets.

In conclusion, the model performance is acceptable in terms of the chosen metrics, and the retrieval of true results. It is not surprising given the extraordinary capacity of hidden Markov Models, optimized and refined in the HMMER tool, when it comes to detecting conserved patterns in symbolic sequences. Since the Kunitz/BPTI domain presents defined and conserved features, as the disulphite bridges with their relative bonding pattern, an HMM is a more then suited statistical model for identifying and detecting its presence, on newly discovered or still unreviewed protein sequences.

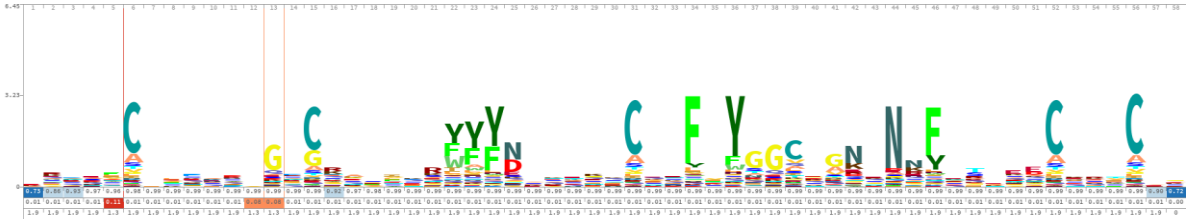

Fig. 3. HMM logo of the BPTI/Kunitz domain generated by Skilign [16].

References

[1] Ranasinghe S, McManus DP. Structure and function of invertebrate Kunitz serine protease inhibitors. Dev Comp Immunol. 2013 Mar;39(3):219-27. doi: 10.1016/j.dci.2012.10.005. Epub 2012 Nov 24. PMID: 23186642.

[2] Kunitz M, Northrop JH. ISOLATION FROM BEEF PANCREAS OF CRYSTALLINE TRYPSINOGEN, TRYPSIN, A TRYPSIN INHIBITOR, AND AN INHIBITOR-TRYPSIN COMPOUND. J Gen Physiol. 1936 Jul 20;19(6):991-1007. doi: 10.1085/jgp.19.6.991. PMID: 19872978; PMCID: PMC2141477.

[3] Laskowski M Jr, Kato I. Protein inhibitors of proteinases. Annu Rev Biochem. 1980;49:593-626. doi: 10.1146/annurev.bi.49.070180.003113. PMID: 6996568.

[4] Thomas E. Creighton, Interactions between cysteine residues as probes of protein conformation: the bisulphide bond between Cys-14 and Cys-38 of the pancreatic trypsin inhibitor, Journal of Molecular Biology, Volume 96, Issue 4, 1975, Pages 767-776, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(75)90151-5.

[5] C.H. Yuan, Q.Y. He, K. Peng, J.B. Diao, L.P. Jiang, X. Tang, S.P. Liang, Discovery of a distinct superfamily of Kunitz-type toxin (KTT) from tarantulas, PLoS One, 3 (2008), p. e3414

[6] de Souza J., Morais K. L.P., Anglés-Cano E., Boufleur P., de Mello E., Augusto Maria D., Taemi Origassa C., Campos Zampolli H., Saraiva Câmara N., Maria Berra C., Viola Bosch R., Chudzinski-Tavassi A. Promising pharmacological profile of a Kunitz-type inhibitor in murine renal cell carcinoma model. Oncotarget. 2016; 7: 62255-62266.

[7] Morjen M, Kallech-Ziri O, Bazaa A, Othman H, Mabrouk K, Zouari-Kessentini R, Sanz L, Calvete JJ, Srairi-Abid N, El Ayeb M, Luis J, Marrakchi N. PIVL, a new serine protease inhibitor from Macrovipera lebetina transmediterranea venom, impairs motility of human glioblastoma cells. Matrix Biol. 2013 Jan;32(1):52-62. doi: 10.1016/j.matbio.2012.11.015. Epub 2012 Dec 20. PMID: 23262217.

[8] Ding Li, Hao Jinbo, Luo Xudong, Chen Zongyun, 2018/08/01 - Engineering varied serine protease inhibitors by converting P1 site of BF9, a weakly active Kunitz-type animal toxin, 120, 10.1016/j.ijbiomac.2018.08.178, International Journal of Biological Macromolecules

[9] Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242.

[10] E. Krissinel and K. Henrick (2004). *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions*. *Acta Cryst.* D**60**, 2256---2268

[11] Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659.

[12] Schrödinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

[13] The UniProt Consortium (2019). UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515.

[14] Credits to: hmmer.org

[15] Stepek, G., McCormack, G., and Page, A. P. (2010). The kunitz domain protein BLI-5 plays a functionally conserved role in cuticle formation in a diverse range of nematodes. Molecular and Biochemical Parasitology, 169(1), 1–11.

[16] Wheeler, T. J., Clements, J., and Finn, R. D. (2014). Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics, 15(1), 7.