An HMM-based method to detect Kunitz domain starting from the BPTI structure

Mario Esposito

MSc in Bioinformatics, University of Bologna

Abstract

Motivation: Bovine pancreatic trypsin inhibitor (BPTI) is one of the most extensively studied globular proteins belonging to the Kunitz-domain family, which is a broad group of protease inhibitors, involved in many biological processes. Likewise, the Kunitz-type proteins are extensively studied for drug development purposes. The aim of this work is to develop an HMM-based method which reliably identifies the presence of the Kunitz domain in UniProtKB/SwissProt sequences.
Results: The HMM-based method, trained on 77 aligned sequences, resulted to be almost a perfect classifier (MCC = 0.997). Furthermore, from this study it is possible to highlight some issues in the UniProtKB/SwissProt annotation policy.
Contact: mario.esposito17@studio.unibo.it
Supplementary information: Supplementary material and code are available at the following GitHub repository: https://github.com/espositomario/HMM_Kunitz

1 Introduction

1.1 Kunitz domains

Kunitz domains 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 (Rawlings et al., 2004). A highly exposed active site residue at position 15 is usually lysine or arginine and it is crucial for the specificity of serine protease inhibition. The architecture and sequence of the bovine pancreatic trypsin inhibitor (BPTI) incorporating the Kunitz domain are shown in **Figure 1**.

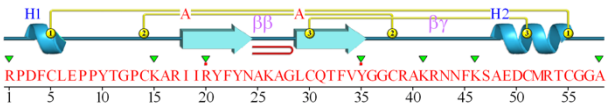

Fig. 1. BPTI architecture and sequence. Three 3 disulfide bridges (yellow), 2 antiparallel β-sheets and helical regions (cyan), 7 phosphate ion binding residues (green) and 2 ligand contact residues (R20,Y35). Source: PDBSum: 1BPI.

Kunitz inhibitors may have a single inhibitory domain or even more forming a multi-domain, single-chain inhibitor. Kunitz peptides exhibit diverse biological activities, such as inhibiting proteases of different classes or adopting new functions such as blocking or modulating ion channels: the Kunitz-type toxin in venomous animals like snakes, spiders, and scorpions (Fry et al., 2009); mammalian inter-alpha-trypsin inhibitors (Fries and Kaczmarczyk, 2003); domain found in Alzheimer's amyloid β-protein in humans (Hynes et al., 1990); domains at the C-termini of the alpha-1 and alpha-3 chains of type VI and type VII collagen (Chen et al., 2013); and tissue factor pathway inhibitor (Jr and Girard, 2012). The research on protease inhibitors has always been in attention owing to their potential applications in medicine, agriculture, and biotechnology (Sabotič and Kos, 2012; Cotabarren et al., 2020).

1.2 Bovine pancreatic trypsin inhibitor (BPTI)

Pancreatic Kunitz inhibitor, also known as aprotinin or BPTI, is one of the most extensively studied globular proteins. It has proved to be a powerful tool for studying protein conformation as well as molecular bases of protein/protein interactions and macromolecular recognition (Ascenzi et al.). BPTI has a relatively broad specificity, inhibiting trypsin as well as chymotrypsin and elastase-like serine enzymes. Clinically, the use of BPTI in selected surgical interventions, such as cardiopulmonary surgery and orthotopic liver transplantation, is advised, as it significantly reduces hemorrhagic complications and thus blood-transfusion requirements (Lemmer et al., 1994; Royston et al., 1987).

1.3 Study workflow

The aim of this study is to develop an HMM-based method which reliably identifies the presence of the Kunitz domain in UniProtKB/SwissProt sequences. In principle, a profile HMM can be derived from unaligned sequences by training. However, the parameters for a profile HMM are more accurately estimated from a multiple sequence alignment (MSA) and this has become the method of choice (Bateman and Haft, 2002). The MSA was retrieved from the alignment of 77 structures similar to the BPTI. The HMM was trained over this MSA and then UniProtKB/SwissProt was adopted to optimize and test the classification performance of the method. The entire workflow of the study is illustrated in **Figure 2**.

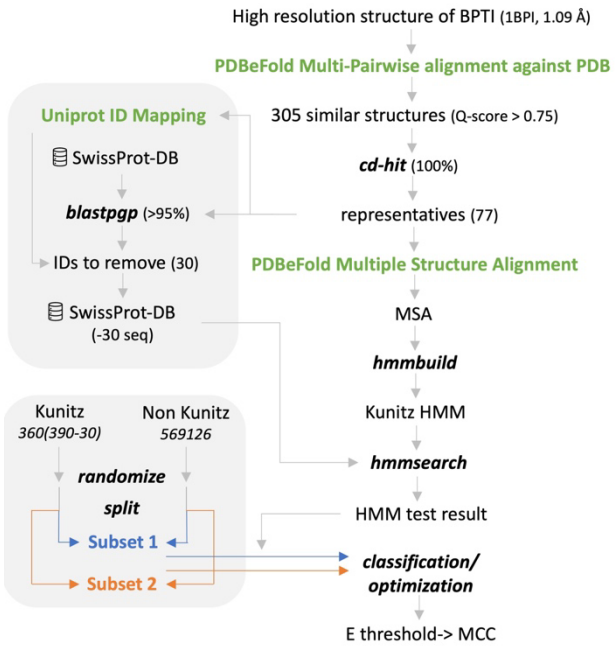

Fig 2. Study workflow. Main pipeline (right), test set preparation (top left) and random generation of the two subsets (bottom left).

2 Methods

2.1 Training set selection

The high-resolution structure of the bovine pancreatic trypsin inhibitor (1BPI) was chosen as the prototype of Kunitz domain to select the seeds (Parkin et al., 1996). The webtool PDBeFold v2.59 was adopted to perform a pairwise structural alignment over the entire PDB (Krissinel and Henrick, 2005; Berman et al., 2000) . PDBeFold was run with default parameters and precision set to 'highest' and the results were selected by a Q-score > 0.75. They were selected according to the Q-score rather than the RMSD because Q-score is computed also considering the alignment length. Q-score **(1)** equation includes the number of matched pairs of C_α atoms (N_{mat}), the number of C_α atoms in the first protein (N_1), the number of C_α atoms in the second protein (N_2) and RMSD.

$$Q_{score} = \frac{N_{mat}^2}{\left[1 + \left(\frac{RMSD}{3}\right)^2\right] \times N_1 \times N_2} \qquad (1)$$

To avoid a bias in the HMM construction, CD-HIT v4.8.1 was adopted to cluster identical sequences and select the longest representative for each cluster (Fu et al., 2012). A threshold of 100% identity was chosen to maintain also sequences differing for only one residue, so that no information on variation went lost.

2.2 MSA and HMM building

A multiple structure alignment between the 77 representativses structures was performed using the PDBeFold v2.59 web. The MSA derived from the structure alignment, was downloaded and adopted as a training set for the HMM training. The hmmbuild program provided by HMMER v3.3.2 was chosen to train the Kunitz's HMM, leaving the optimal trimming of the MSA to the algorithm (Eddy, 2011). The HMM profile logo, displayed in **Figure 3**, was plotted with Skylign (Wheeler et al., 2014).

2.3 Test set preparation

The entire UniProt/SwissProt release_2023_02 (SP) was chosen as a test set and the annotation of Kunitz domain (PF00014) according to PFAM v35.0 was chosen as reference to evaluate the classification performance. Before testing the model, the test set was elaborated in order to avoid a bias in the model evaluation. The seed sequences and all high similarity proteins (>95%), were removed from SP in order to perform a fair test of the HMM. To identify the high -similarity proteins, blastpgp v2.2.26 (gapped-BLAST) was run with default parameters, using the 77 training sequences as queries and the entire SP as target database (Altschul et al., 1997). After that, all the sequences in SP were tested with hmmsearch using the option '--max' which excludes all the heuristic filters. By default, hmmsearch reported in the result only sequences over an E-value threshold of 10. Since the computation of the E-value is influenced by the database search space (Finn et al., 2011), the test was performed once on the entire SP and then the random splitting was applied in order to avoid a bias due to different database size.

2.4 E-value optimization and classification benchmark

In order to select the E-value maximizing the classification performance, the entire SP ID list was split into 2 subsets, using a python script. The subsets lists were generated randomly of equal length and with the same proportion of Kunitz proteins. Subset1 was adopted to select the best threshold and subset2 was adopted as the test set. The role of the 2 subsets was then swapped to cross-validate the results. The 2 subsets lists were annotated with either 1 or 0 depending on the presence of the Kunitz domain according to PFAM and with the E-value previously resulted from hmmsearch. Since the distribution of Kunitz and non-Kunitz was skewed, Matthews's correlation coefficient (MCC) **(2)** was adopted as classification score. Compared to accuracy **(3)** or F1 score, the MCC is a more reliable statistical coefficient which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset (Chicco and Jurman, 2020; Matthews, 1975).

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

$$ACC = \frac{TP \times TN}{TP + TN + FP + FN} \qquad (3)$$

The classification benchmark was tested by running a python script (Supplementary material) for an E-value threshold decreasing exponentially from 1e-1 to 1e-12. For each subset, the E-value threshold for which the model guaranteed the best MCC was identified, and after that, it was verified that the same outcome was achieved for the other subset. The average between the best threshold for both subset was applied in benchmarking the classification for the entire SP. The entire pipeline was runned on MacOS 13.3.1 with python 3.10.

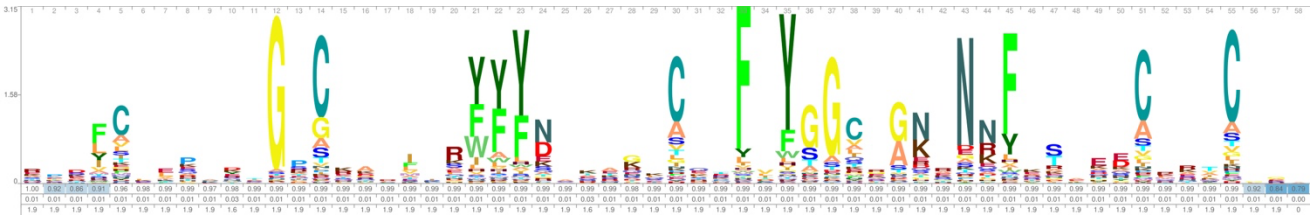

Fig 3. HMM profile logo. It shows clearly the 6 conserved cysteines involved in the 3 disulfide bond, and from position 33 to 51 the BPTI motif (F*Y*GC****N*F*****C).

3 Results

3.1 Training sequence selection and HMM building

The selected BPTI structure (1BPI) is long 58 aa, was refined with a resolution of 1.09 Å, maps from position 36 to position 93 of the entire sequence and comprises the entire Kunitz domain PF00014 (Parkin et al., 1996). This structure was the query of the similarity search with the entire PDB, which resulted in 305 structures with a Q-score > 0.75 (Supplementary material). Structures with identical sequences were grouped together and the longest sequence was chosen as representatives for each of the 77 clusters. Multiple structure alignment between the 77 representatives resulted in an overall RMSD = 0.9503, overall Q-score = 0.6582 and 49 residues aligned (Supplementary material). The MSA derived from the structure alignment was adopted to train the HMM, which resulted with a length of 58 as it is displayed in **Figure 3**.

3.2 Test set cleaning

Thirty sequences were excluded from SP, in order to do a fair classification performance test. Twentynine sequences had more than 95% of identity with at least one of the seeds (Supplementary material). To check whether blastpgp was able to identify all the training sequences, the 77 PDB IDs were mapped to SP with UniProt ID Mapping, resulting in a match with 16 UniProt IDs. Six PDB IDs were not mapped in SP (3CI7:A, 6HAR:E, 3L3T:E, 3AUE:C, 5NX3:C and 3AUC:A). 6HAR:E was mapped to UniProt/TrEMBL while the other PDB IDs belong to protein variants. Blastpgp was able in identifying all the training sequences apart from one sequence corresponding to the PDB 5M4V:A, mapping the UniProt ID A0A1Z0YU59. The sequence on PDB differs in 3/57 residues compared to the UniProt sequence. For this reason, the alignment in blastpgp resulted in an identity of 94.7% (54/57), and the threshold of 95% was not able to filter in this entry. However, it was strange to notice that no mutations were reported on PDB. So, overall the 16 training sequences and 14 high similarity sequences were removed from SP (#seq = 569516-30 = 569486). As expected, these sequences were all annotated with a Kunitz domain according to PFAM.

3.3 Classification benchmark

To ensure the E-value threshold optimization independent from the test set chosen, a kind of 2-fold cross validation was performed. Therefore, SP was split in two halves so that each subset comprised 284743 proteins of which 180 were Kunitz proteins randomly distributed. Although the E-values were assigned to the subsets, they derived from the HMM test performed on the entire SP and thus they have to be evaluated with respect to a database size of 569486.

Since the distribution of the Kunitz class was skewed, MCC was chosen as a reliable statistical coefficient to assess the classification performance. For each subset the MCCs corresponding to an E-value threshold decreasing exponentially from 1e-1 to 1e-12 were computed. For each subset the highest E-value threshold generating the best MCC was selected. The E-value threshold of 1e-3 produced the best MCC of 0.997 in the subset1 classification. On the other hand, the E-value threshold ranging between 1e-7 and 1e-3 produces the best MCC of 0.997 in the subset2 classification. The line plot in **Figure 4** shows the different MCCs obtained over the E-value range for both subsets.

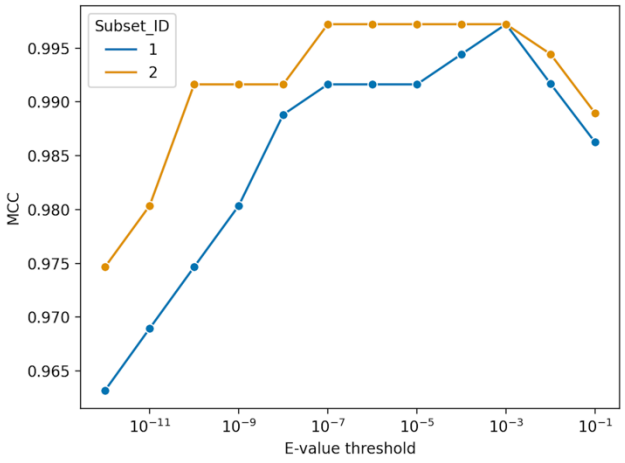

Fig. 4 E-value threshold vs MCC. It is displayed the relationship between the E-value threshold (x-axis) and the corresponding Matthews Correlation Coefficients (MCCs) obtained after classification (y-axis). Both subset1 (blue) and subset2 (orange) are displayed.

To be lenient, within the optimal range, the highest value 1e-3 was selected as the optimal threshold for subset2. Since the optimal threshold was the same for both subsets (1e-3), it was applied to test the whole SP. The classification outcome on the whole SP resulted in a confusion matrix containing: 358 TP, 569126 TN, 2 FN and 0 FP. The MCC was of 0.997 and the 2 FN were associated with the UniProt entries: D3GGZ8 and O62247 (Supplementary material).

Discussion and conclusion

Firstly, the results indicate that this HMM model achieved a good classification performance when tested on the entire SP. The two IDs misclassified as false negatives seem to be not correctly annotated by PFAM. On the UniProt they are associated with BLI-5 protein of two different nematodes (*Haemonchus contortus* and *Caenorhabditis elegans*) and both have a Kunitz domain inferred by PFAM. However, there is a warning label on the function that states : 'Appears to have serine protease activity in vitro.' Indeed, in vitro assays suggest that recombinant BLI-5 in nematodes are proteolytic enzymes and not inhibitors of serine protease activity (Stepek et al., 2010). Furthermore, it is worth noting that the BPTI motif (**Figure 3**) is missing in the sequences of the BLI-5, suggesting the absence of the inhibition function. After all, one observation can be: since, PFAM does not rely on experimental evidence but on the statistical property of HMMs, it's possible to individuate some misclassifications in which an annotated domain does not show the expected function in experimental assays. Given this observation, this model, which follows the same principles as PFAM and is optimized on PFAM itself, may also have the same vulnerabilities.

Overall, the satisfactory performance of the HMM model on the entire SP dataset indicates its effectiveness in classifying a wide range of proteins. This suggests that the model has the potential to be applied to large-scale protein annotation efforts, where accurate and efficient classification is essential. On the other hand, the misclassification of certain protein domains highlights the limitations of relying solely on statistical properties for function prediction.

Moreover, combining the HMM model with other computational methods can further enhance our understanding of protein function. Machine learning algorithms, for example, can be employed to extract additional features from protein sequences or to integrate diverse data sources. By leveraging the strengths of different approaches, we can achieve a more comprehensive and accurate prediction of protein function.

References

Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

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

Bateman,A. and Haft,D.H. (2002) HMM-based databases in InterPro. Brief. Bioinform., 3, 236–245.

Berman,H.M. et al. (2000) The Protein Data Bank. Nucleic Acids Res., 28, 235–242.

Chen,P. et al. (2013) Collagen VI in cancer and its biological mechanisms. Trends Mol. Med., 19, 410–417.

Chicco,D. and Jurman,G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6.

Cotabarren,J. et al. (2020) Biotechnological, biomedical, and agronomical applications of plant protease inhibitors with high stability: A systematic review. Plant Sci., 292, 110398.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. PLOS Comput. Biol., 7, e1002195.

Finn,R.D. et al. (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res., 39, W29–W37.

Fries,E. and Kaczmarczyk,A. (2003) Inter-alpha-inhibitor, hyaluronan and inflammation. Acta Biochim. Pol., 50, 735–742.

Fry,B.G. et al. (2009) The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. Annu. Rev. Genomics Hum. Genet., 10, 483–511.

Fu,L. et al. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28, 3150–3152.

Hynes,T.R. et al. (1990) X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid .beta.-protein precursor. Biochemistry, 29, 10018–10022.

Jr,G.J.B. and Girard,T.J. (2012) Tissue factor pathway inhibitor: structure-function. Front. Biosci.-Landmark, 17, 262–280.

Krissinel,E. and Henrick,K. (2005) Multiple Alignment of Protein Structures in Three Dimensions. In, R. Berthold,M. et al. (eds), Computational Life Sciences, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 67–78.

Lemmer,J.H. et al. (1994) Aprotinin for coronary bypass operations: Efficacy, safety, and influence on early saphenous vein graft patency: A multicenter, randomized, double-blind, placebo-controlled study. J. Thorac. Cardiovasc. Surg., 107, 543–553.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA - Protein Struct., 405, 442–451.

Parkin,S. et al. (1996) Structure of bovine pancreatic trypsin inhibitor at 125 K definition of carboxyl-terminal residues Gly57 and Ala58. Acta Crystallogr. D Biol. Crystallogr., 52, 18–29.

RAWLINGS,N.D. et al. (2004) Evolutionary families of peptidase inhibitors. Biochem. J., 378, 705–716.

Royston,D. et al. (1987) EFFECT OF APROTININ ON NEED FOR BLOOD TRANSFUSION AFTER REPEAT OPEN-HEART SURGERY. The Lancet, 330, 1289–1291.

Sabotič,J. and Kos,J. (2012) Microbial and fungal protease inhibitors—current and potential applications. Appl. Microbiol. Biotechnol., 93, 1351–1375.

Stepek,G. et al. (2010) 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.

Wheeler,T.J. et al. (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics, 15, 7.