Hidden Markov Model-Based Annotation of Kunitz Domains: Development and Performance Evaluation

Irene D'Onofrio[1]

[1]Università di Bologna, MSc Bioinformatics

Abstract

Motivation: Kunitz-type serine protease inhibitors are found in several organisms including animals, plants, and microbes. In the recent years the Kunitz domain has gained significant attention in the field of protein engineering, as it holds great potential for the development of specific protease inhibitors with significant relevance in therapeutic applications. Given its relevance, it seems crucial to develop and ameliorate tools for its annotation. Thus, the aim of this study was to develop an HMM-based approach that could efficiently annotate this domain within a broad range of proteins.

Results: The Hidden Markov Model, built from 30 structurally well-defined proteins, resulted to be an efficient classifier (MCC = 0.997).

Contact: irene.donofrio@studio.unibo.it

Supplementary information: Supplementary material and code are available at the following GitHub repository: https://github.com/DOnofrio-Irene/Kunitz_HMM_prj

1 Introduction

Kunitz protease inhibitors, whose chief member is the well-known pancreatic trypsin inhibitor (BPTI), are ubiquitous proteins, found spanning the evolutionary tree from microbes to mammals, as serine protease inhibitors or animal toxins in venomous animals. Kunitz-type inhibitors are classified into two families, namely I2 and I3, as categorized by MEROPS database (Rawlings et al., 2004b). The I2 family comprises Kunitz inhibitors found in animals, which specifically inhibit proteases belonging to the S1 family, primarily serine proteases. On the other hand, the I3 family consists of Kunitz inhibitors derived from plants, which predominantly target serine proteases but also exhibit inhibitory activity against aspartic proteases (A1) and cysteine proteases (C1) (Rawlings et al., 2004a). These proteins can have single or multiple Kunitz inhibitory domains linked or associated with other domain types. Kunitz domains, ranging from 3 to 20 kDa in size, 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. The comparison of the sequences of various Kunitz-type domains reveals the presence of six cysteine residues that exhibit high conservation (Fig.1). This observation aligns with crystallographic studies, which confirm the existence of three disulfide bridges responsible for stabilizing the protein's structural integrity. This arrangement is characterized by the bonding patterns C1–C6, C2–C4, and C3–C5 . Two of the disulfide bonds (C1–C6 and C3–C5) are required for the maintenance of native conformation, whereas the third (C2–C4) stabilizes the two binding domains (Laskowski and Kato, 1980).

The P1 position, located at residue 15, plays a crucial role in determining the specificity of serine protease inhibition. Typically, this position is occupied by either arginine (Arg) or lysine (Lys), which are highly exposed and inserted into the S1 site of the corresponding protease. The high abundance of basic amino acids and the presence of the guanidine moiety on the side chain of arginine are the reason for the overall high isoelectric point. Kunitz-type domains exhibit diverse forms, conferring specific serine protease inhibitory functions to associated proteins. These forms include the Kunitz-type toxin (KTTs) found in venomous animals such as snakes, spiders, and scorpions, the mammalian inter-alpha-trypsin inhibitors, the domain present in Alzheimer's amyloid β-protein in humans domains located at the C-termini of the alpha-1 and alpha-3 chains of type VI and type VII collagen, and the tissue factor pathway inhibitor (TFPI) (Mishra, 2020).

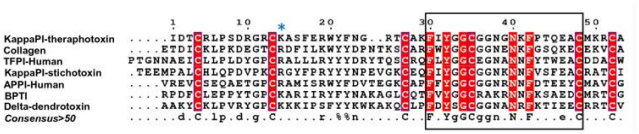

Fig. 1. Amino acid sequence alignment of Kunitz-domain inhibitors from different source organisms. Conserved residues are highlighted in red and the P1 residue (K/R) is marked with an asterisk. Kunitz family signature sequence (F*Y*GC****N*F*****C) is shown in a box.

Phylogenetic analyses showed that the ancestral gene of the Kunitz-type inhibitor appeared about 500 million years ago. Thereafter, this gene duplicated itself many times, and some of the duplicates were inserted into other protein-coding genes (Ikeo et al., 1992). A recent study found strong evidence for positive selection acting on the snake Kunitz/BPTI genes, demonstrating that adaptive evolution was a force driving their evolution. The main evolutionary pressure driving these variations could be the necessity of functional diversity against the host molecular targets which also keep co-evolving rapidly during the arms race (Župunski and Kordiš, 2016). The dual functionality (serine pro-

tease inhibitors and ion channel blockers) and properties exhibited by Kunitz-type toxins (KTTs) have generated significant interest, not only in the field of evolutionary research but also as potential candidates for pharmacological investigations (Thakur and Mukherjee, 2017). Indeed, by targeting proteases implicated in pathological conditions, such as cancer, inflammation, and neurodegenerative disorders, Kunitz protein inhibitors hold the potential to develop targeted therapies with improved efficacy and reduced side effects (Mukherjee and Mackessy, 2014; de Souza *et al.*, 2016).

Given their clinical relevance, the development and improvement of tools for annotating Kunitz domains are crucial. Various approaches have been devised to define and identify protein domains, with some relying on a structural classification system and others inferring domains through clustering conserved subsequences. Hidden Markov Models (HMMs) have proved to be a powerful tool for this task (Yoon, 2009). Here, it's described the process of generating a Hidden Markov Model capable of annotating Kunitz domains within the UniProtKB/Swiss-Prot database.

2 Methods

2.1 Training set selection

The selection of a representative training set was carried out by means of an advanced search in the RCSB PDB database (wwPDB consortium, 2019), followed by a clustering procedure performed with CD-HIT version 4.8.1 (Fu *et al.*, 2012). The RCSB PDB database was queried to fetch a collection of structures with annotated Kunitz domain by Pfam (Pfam identifier: PF00014), refinement resolution below 2.50 Å, and polymer entity sequence length between 50 and 90 residues. To eliminate the redundancy of the PDB structures, so not to bias the HMM generation, CD-HIT was adopted using a sequence identity threshold of 0.95. CD-HIT is a greedy incremental algorithm that starts with the longest input sequence as the first cluster representative and then processes the remaining sequences from long to short to classify each sequence as a redundant or representative sequence based on its similarities to the existing representatives.

2.2 MSA and HMM generation

The representative list was submitted to PDBeFold v2.59 alignment program to obtain a multiple structure alignment (Supplementary Material) (Krissinel and Henrick, 2004). The resulting seed MSA was provided as input to hmmbuild function of HMMER 3.3.2. to obtain a profile-HMM (Supplementary material) (Finn et al., 2011). To visualize the constructed HMM, the Skylign webserver was employed (Wheeler et al., 2014).

2.3 Selection of the test set

In order to test and validate the model, a dataset comprising both proteins containing the Kunitz domain and those not classified as such was necessary. Hence, the entire UniProtKB/Swiss-Prot database, containing 569516 sequence entries (release 2023_02 of 03-May-2023), was downloaded. To ensure a fair evaluation of the HMM model, proteins sharing a high level of sequence identity with the representatives were excluded from the test set. Identification of redundant proteins was carried out using the blastpgp program, a specialized protein BLAST comparison tool that offers increased sensitivity compared to standard BLASTP searches (Altschul et al., 1997). The blastpgp command was run with default parameters, using as queries the FASTA file containing the representative sequences and the entire UniProtKB/Swiss-Prot database (results available in the Supplementary material). Sequences showing more than 95% of sequence identity were considered redundant and compiled into a list, provided in the Supplementary Material. Additionally, to ensure the inclusion of all training sequences, a comparison was performed between the proteins identified with blastpgp and the UniprotKB IDs of the representatives. To remove the redundant sequences from the test set a Python script (`rem_fasta_seqs.py` in Supplementary Materials) was used.

2.4 Model testing

To account for the influence of database size on the E-values, the `hmmsearch` command from the HMMER software was run against the entire test set, with the option '--max' which excludes all the heuristic filters.

The `hmmsearch` command reads an HMM from hmmfile and searches seqfile for significantly similar sequence matches. The output generated by hmmsearch was subsequently utilized in the `subsets-creation.py` Python script (provided in the Supplementary Materials). The `subsets-creation.py` script was designed to generate two subsets of equal size, with the same representation of both Kunitz and non-Kunitz proteins (subset1 and subset2, Supplementary Materials).

Each protein in the subsets is associated with its corresponding e-value (obtained with the hmmsearch program) and label (0 or 1 based on the absence or presence of the Kunitz domain, respectively). The labeling process and the reintroduction of those proteins which weren't shown in the hmmsearch were performed by the Python script with a comparison between the results and the lists of Kunitz and non-Kunitz proteins. The lists were downloaded from Uniprot by performing an advanced search: the first query aimed to identify Kunitz proteins, and the constraint imposed was that the entries must possess the Pfam identifier pf0014. Subsequently, a second query was executed to identify non-Kunitz proteins, and the constraint was set such that the entries should not have the Pfam identifier PF00014. For the proteins not identified by hmmsearch, a fictional E-value of 999 was assigned to ensure their inclusion in subsequent analyses.

2.5 E-value optimization and performance measurement

In order to pick the optimal E-value, able to maximize the classification performance, an optimization procedure was carried out on the two subsets derived from the random splitting of the whole test set. The performance was tested by executing the `performance.py` script (Supplementary material).

In the optimization the script was executed on a subset for a range of E-values (1e-1 - 1e-12); the threshold that yields the best metrics (best MCC) was selected and adopted for the testing on the opposite subset, to verify whether the outcome was simi

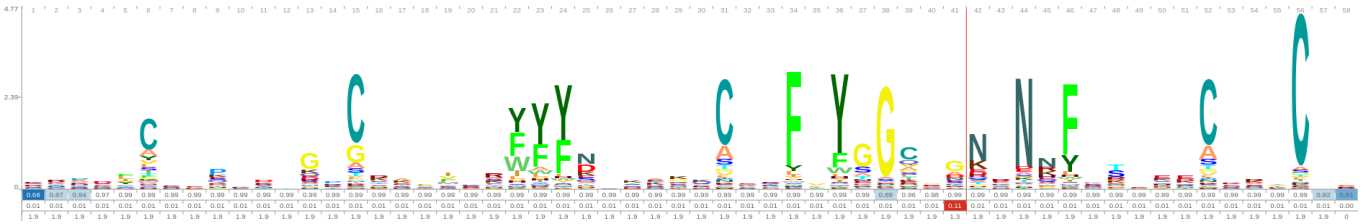

Fig.2 HMM profile logo: the profile logo generated with Skylign clearly highlights the six conserved cysteine residues.

lar. The role of the two subsets was then swapped. The results of the optimization subsets for each E-value are available in the Supplementary Material. To evaluate the performance of the model, the average of the two optimal E-values was tested on the entire test set.

In these performance measurements the Matthews correlation coefficient (MCC) was adopted to evaluate the efficacy of the model (Fig 2). The choice is justified by the unbalanced nature of the test set. MCC is a measure unaffected by the unbalanced datasets issue, indeed, with MCC to get a high-quality score, the classifier has to make correct predictions both on the majority of the negative cases, and on the majority of the positive cases, independently of their ratios in the overall dataset. F1 and accuracy (2), instead, generate reliable results only when applied to balanced datasets, and produce misleading results when applied to imbalanced cases (Chicco and Jurman, 2020).

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

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

3 Results

3.1 Training set selection, MSA and HMM generation

The advanced search on PDB resulted in the recovery of 141 polymer entities. Structures sharing more than 95% of sequence identity were clustered together. The list of representatives counted 30 entities (Supplementary Materials). The multiple structure alignment resulted in 39 aligned residues, 3 aligned SSEs, an overall RMSD of 0.7623, and an overall Q-score of 0.5400. HMMER turned the multiple structure alignment into a profile of 58 consensus positions, which can be found in the Supplementary material. The HMM logo is showed in Fig.2.

3.2 Test set generation

The search performed by blastpgp resulted in 33 proteins that shared a sequence identity greater than or equal to 95% with at least one of the representatives. Although belonging to the training set, Mambaquaretin-1 (PDB ID: 5M4V_1, UniProtKB ID: A0A1Z0YU59) was not initially included in the redundant proteins list. The Mambaquaretin-1 protein exhibited a sequence identity of 94.737%, since on an alignment length of 57 residues there are 3 mismatches. Indeed, looking at its PDB page, it is possible to visualize the differences between the 5M4V_1 sequence and the UniProt corresponding sequence in positions 15,16, and 48.

To rectify this omission, the Mambaquaretin-1 protein was manually added to the list of redundant proteins.

Ultimately, 34 sequences were removed from the UniProtKB/Swiss-Prot database (test set).

3.3 Classification and performance results

The optimization procedure was carried out on the two equally partitioned subsets, each containing 284741 entries. Since the hmmsearch was executed on the whole test set they derived from, it was possible to compare the E-values between them.

Table 1. Optimization results of threshold

	SET	threshold	MCC	ACC
OPTIMIZATION	Subset1	0.001	1.00	1.00
TEST	Subset2	0.001	0.994	0.999
OPTIMIZATION	Subset2	0.001	0.994	0.999
TEST	Subset1	0.001	1.00	1.00
FINAL TEST	Test set	0.001	0.997	0.999

The optimization on the subset 1 yielded as best threshold 0.001, with an MCC of 1, Similarly, for optimization on subset 2, an E-value of 0.001 also produced the best MCC of 0.9943 (Table 1). Notably, when examining the MCC vs E-value graph (Fig. 3a), a similar trend of MCC can be observed for both subsets within the E-value range of 1e-2 to 1e-7. As the best threshold remained consistent for both optimizations, the E-value of 0.001 was selected to evaluate the Hidden Markov Model on the entire test set, which comprised 569,482 entries.

The results of the classification resulted in an MCC of 0.997 and a confusion matrix with 354 true positives, 0 false positives, 2 false negatives and 569126 true negatives, as shown in Fig 3b. The two false negatives are proteins associated with UniProt entries O62247 and D3GGZ8 respectively.

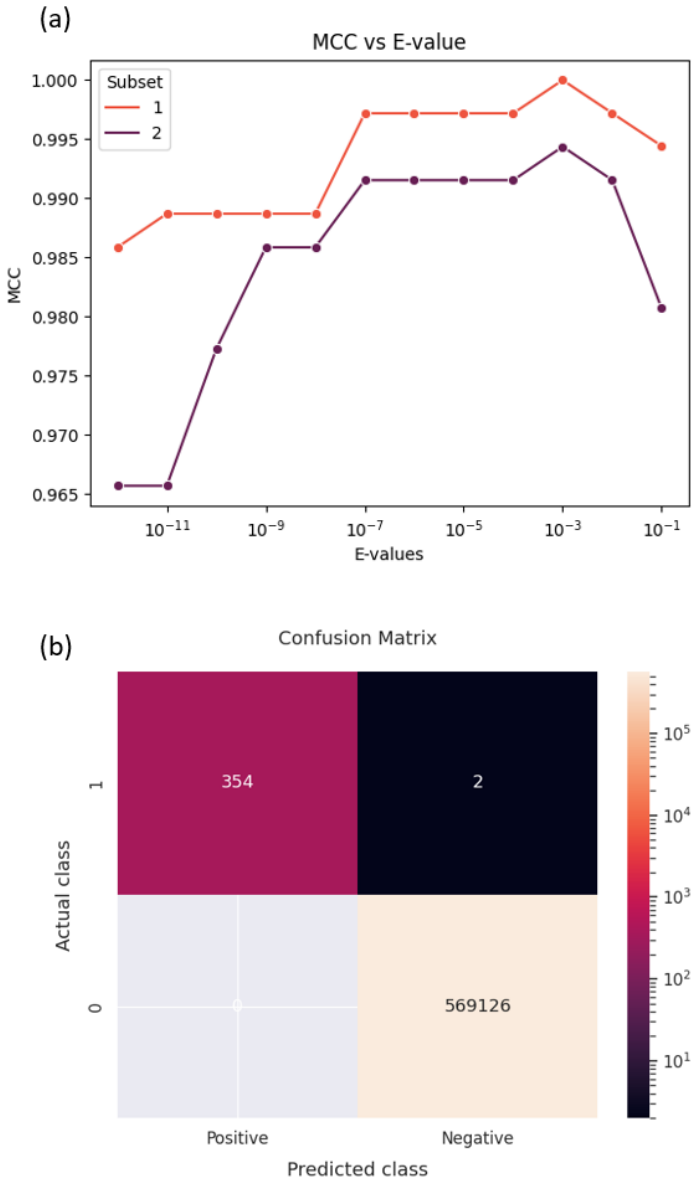

Fig.3. a) Relationship between E-value and MCC in the optimization results: It is possible to visualize the similar trend for the two subsets in the range $1e^{-9}$-$1e^{-3}$**; b) Confusion matrix of the final results.**

4 Discussion and conclusion

The objective of this study was to construct a Hidden Markov Model that could effectively annotate the Kunitz domain, starting from a set of structurally defined representative proteins. Based on the obtained performance outcomes, it is possible to state that this HMM succeeded in the classification task when tested on a dataset of >500000 entries.

Nonetheless, the Kunitz-type proteins BLI-5 from *Caenorhabditis elegans* (UniProt ID O62247) and *Haemonchus contortus* (Barber pole worm) (UniProt ID D3GGZ8) were both misclassified and labelled as false negatives, with E-values of 0.046 and 4.8, respectively. Upon examining the alignment between the model and the domain (see Supplementary material), it becomes apparent that not only the sequence identity is very low, but also that both proteins lack one of the cysteine residues (C4). As a result, one of the characteristic three disulfide bonds is missing, as evident from the fact that only two bonds are indicated in the 'PTM/Processing' subsection of these two entries. However, as aforementioned, the C2-C4 is not responsible for the maintenance of the native conformation (Laskowski and Kato, 1980). As it has been demonstrated by experimental results, BLI-5 proteins do not inhibit the serine protease activity, but instead, they act as proteolytic enzymes. This behaviour can be explained by the absence of key residues from the bovine pancreatic trypsin inhibitor motif ($Fx_{(3)}GCx_{(6)}FYx_{(5)}C$), indeed, the identity between Kunitz domain of BLI-5 protein from each of the nematode species and bovine pancreatic trypsin inhibitor is only approximately 20% (Stepek et al., 2010). It is also worth noticing that these two entries are classified as false negatives also by the PROSITE entry PS50279, the predictor associated with the UniRule PROSITE-ProRule PRU00031, used to annotate the BPTI/Kunitz inhibitor. Despite the overall effectiveness of the model in classifying a wide range of proteins, it is important to note that being solely based on probabilistic properties, it is susceptible to errors, as demonstrated by this study.

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.

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.

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

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

Ikeo,K. et al. (1992) Evolutionary origin of a Kunitz-type trypsin inhibitor domain inserted in the amyloid beta precursor protein of Alzheimer's disease. J. Mol. Evol., 34, 536–543.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr., 60, 2256–2268.

Laskowski,M. and Kato,I. (1980) Protein inhibitors of proteinases. Annu. Rev. Biochem., 49, 593–626.

Mishra,M. (2020) Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. J. Mol. Evol., 88, 537–548.

Mukherjee,A.K. and Mackessy,S.P. (2014) Pharmacological properties and pathophysiological significance of a Kunitz-type protease inhibitor (Rusvikunin-II) and its protein complex (Rusvikunin complex) purified from Daboia russelii russelii venom. Toxicon Off. J. Int. Soc. Toxinology, 89, 55–66.

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

Rawlings,N.D. et al. (2004b) MEROPS: the peptidase database. Nucleic Acids Res., 32, D160-164.

de Souza,J.G. et al. (2016) Promising pharmacological profile of a Kunitz-type inhibitor in murine renal cell carcinoma model. Oncotarget, 7, 62255–62266.

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.

Thakur,R. and Mukherjee,A.K. (2017) Pathophysiological significance and therapeutic applications of snake venom protease inhibitors. Toxicon, 131, 37–47.

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.

wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res., 47, D520–D528.

Yoon,B.-J. (2009) Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr. Genomics, 10, 402–415.

Župunski,V. and Kordiš,D. (2016) Strong and widespread action of site-specific positive selection in the snake venom Kunitz/BPTI protein family. Sci. Rep., 6, 37054.