Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Giacomo Orsini[1,*]

[1]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[*]Corresponding author. giacomo.orsini2@studio.unibo.it

Abstract

Motivation: Kunitz-type domains represent one of the broader groups of protease inhibitors. Their functions are various, and their presence is almost ubiquitarian in the evolutionary tree. Recently, the interest in this family of proteins has arisen, given their potential applications in medicine and agriculture. The structure of this domains is characterized by six conserved Cysteine residues, which are responsible for their structural conformation. The aim of this work is to develop an HMM-based method that can reliably identify the presence of Kunitz domain(s) in UniProtKB/SwissProt protein sequences.

Results: The results have found that, among four profile HMMs trained on different sets of Kunitz domain structures, the ones built without mutated domains perform better; moreover, by modifying the length of the considered structures, results have shown that the models trained on structures with a low and middle range of lengths perform better until an E-value threshold of 1E-05, while the model built on a broader range performs better at a threshold of 1E-07, with lower performances.

Supplementary information: supplementary data are available at `https://github.com/giacomoorsini/profile-HMM_project`

Introduction

Peptidases are crucial for the survival of all living organisms, as, by degarading proteins, they make amino acids accessible to biological functions. Their activity must be regularized by strict pathway-regulation mechanisms, which often include the use of protease inhibitors. There are many classes of protease inhibitors, divided by mechanism and structural similarity (16). Among these, one of the broader groups is composed by the Kunitz-type protein domains. While their main function is the inhibition of serine proteases (19) (GO:0004867 (10)), these inhibitors can perform a vast range of functions and are ubiquitarian, as they can be traced along all the evolutionary tree, except in the virus *Amsacta moorei entomopoxvirus* (8).

Kunitz-domain inhibitors that are only present in animals are classified under the inhibitor family I2, Clan IB (2). There are at least three main topologies of Kunitz-type inhibitors (21): proteins may contain one single Kunitz domain (e.g.: bovine pancreatic trypsin inhibitor(BPTI)-like proteins), two Kunitz domains (e.g.: urinary trypsin inhibitor (UTI) and hepatocyte growth factor activator inhibitor (HAI)) or even three tandemly arranged Kunitz domains (e.g.: tissue factor pathway inhibitor (TFPI)-like molecules) (Figure 1).

In mammals, they can be found in the plasma, saliva, stomach, and several other tissues. Moreover, these domains are often present as an insertion in various proteins, most likely due to exon shuffling, thereby incorporating domains from different evolutionary precursors (15).

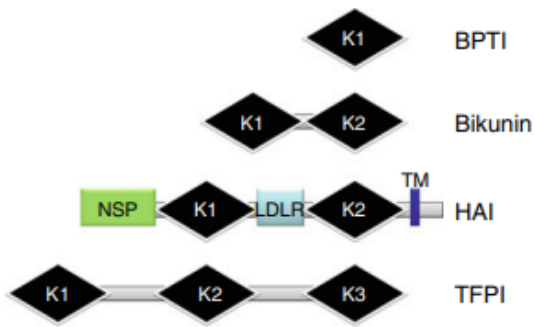

Fig. 1.: Main topologies of Kunitz like proteins. The Bikunin topology is typical of protease inhibitors which have two tandemly arranged Kunitz domains (21).

Functionally, Kunitz-domains are known to be involved in various physiological processes such as host defense against microbial infection, blood coagulation, fibrinolysis, and inflammation (16). Because of this, medical research on these protease inhibitors has always been under the spotlight (14). They are also known as frequent components of venom from poisonous animals such as snakes, spiders, and even ticks, acting as ion channel blockers and anticoagulants.

Fig. 2.: Bovine Pancreatic Trypsin inhibitor 1BPT from PDB(7). Higlightened, the 6 Cysteines responsible for the folding pattern, according to the 1-6,2-4,3-5 binding pattern: pink Cys5-Cys55, red Cys30-Cys51, yellow Cys14-Cys38. For the visualization, we used PDB viewer tool.

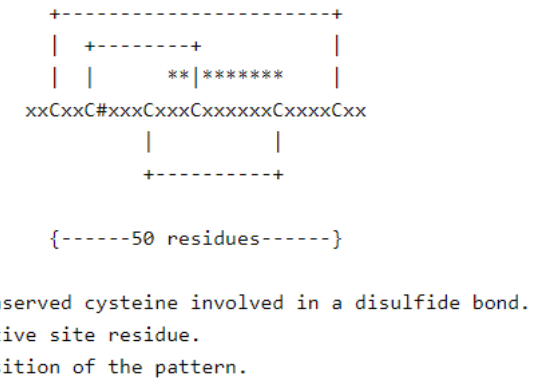

Fig. 3.: Cysteine binding pattern of the Kunitz-type domains according to PFAM.

One of the most studied Kunitz-type protease inhibitors, often used as a model structure, is the Bovine pancreatic trypsin inhibitor (BPTI or basic protease inhibitor), also known as aprotinin, and commercialized as an anticoagulant drug (Trasylol). Other examples of notorious Kunitz inhibitors are the Alzheimer's amyloid precursor β-protein (APP), tissue factor pathway inhibitor (TFPI), domains at the C-termini in type VII and type VI collagens and the Kunitz-soybean trypsin inhibitor (STI), a protein belonging to the inhibitor family I3A (2), which was the first Kunitz protein to be studied (22).

Kunitz domains from the I2 family can function as standalone protease inhibitors in their free form by recognizing specific protein structures (16), in particular serine-type endopeptidases. They are typically 50-80 amino acids long, have a weight of 6 kDa, a globular shape and a basic behavior given by the high amount of positively charged residues. Their conserved structural fold is composed of two antiparallel β-sheets and one or two α helical regions that are stabilized with three disulfide bridges (bonding pattern of 1–6, 2–4, 3–5, as seen in Figures 2 and 3) (8). These six Cysteine residues are therefore conserved, as they also present the protease-binding loop at its surface. This active site, named P1, usually has a Lysine or an Arginine in position 15, which is responsible for the specificity of the bond with an Aspartate in position 189 of the S1 domain of the serine protease. Based on the amino acid residues at the functional site of various Kunitz-type inhibitors, it is inferred that this 'flexibility within the structural rigidity' is responsible for the functional heterogeneity of Kunitz-domain inhibitors (16).

The Pfam database (9) groups within the Kunitz BPTI domain family (PF00014, IPR002223 on InterPro (8)) more than 30,000 sequences, 390 of which have been manually reviewed (available on UniProt/SwissProt database (12)), and 194 PDB structures organized in 2475 possible architectures.

A hidden Markov model (HMM) is a statistical model that can be used to describe the evolution of observable events that depend on not directly observable internal factors. Profile-HMMs have been shown to be very effective in modelling biological sequences, protein classification, motif detection etc., due to the convenience and effectiveness in representing sequence profiles (23).

Aim of the work

The aim of this work was to train and test an HMM profile capable of correctly annotating new amino acid sequences as Kunitz domains. Kunitz domains protein structures retrieved from the RCSB PDB (11) database were used to train the models, which were then tested on protein sequences found in the Uniprot/Swissprot database. The workflow was split into two rounds: during the first round, 2 models were trained and tested to retrieve the best performance; in the second round, the same pipeline was used to create and test 2 additional models starting from the features of the best of round 1. In the end, the performance was calculated and the best HMM profile between the 3 was selected. Accuracy and Matthew's correlation coefficient (MCC (13)) measurements, as well as F1 score, were used as performance evaluators.

Materials and methods

As mentioned, our workflow can be divided into two rounds, since the evaluation of the performance has been conducted two times. The pipeline that is hereby reported has been used to create each one of the 4 HMM profiles.

Databases

The employed databases have been:

- RCSB PDB (as of 05/2023) (11) to retrieve protein structures for building the HMM profiles.
- UniProtKB/SwissProt (as of 05/2023) (12) to download sequences of proteins both containing and not containing a Kunitz domain, according to the Pfam annotation.

Training sets

The first step of the pipeline has been the development of a training set to create the HMM profile. The retrieval of the Kunitz domains structures has been conducted with the use of the advanced search

Table 1. PDB queries

Subquery	Round 1		Round 2	
	model 1	model 2	model 3	model 4
PFAM identifier:	PF00014	PF00014	PF00014	PF00014
CATH indentifier:	4.10.410.10	4.10.410.10	4.10.410.10	4.10.410.10
SCOP identifier:	4003337	4003337	4003337	4003337
Resolution:	<=3	<=3	<=3	<=3
Length range:	51-76	51-76	50-60	50-80
Structure title has	mutant, mutagenesis, mutation		mutant, mutagenesis, mutation	mutant, mutagenesis, mutation
not any of words:	BPTI-mutant, variant, Variants		BPTI-mutant, variant, Variants	BPTI-mutant, variant, Variants
Mutations count:	0		0	0

Subqueries of the PDB advanced search for every model.

on RCSB PDB database. The selected specifications for each model are shown in table 1.

We have chosen structures that were annotated as Kunitz-type domains in 3 major classification databases: SCOP (20), Pfam (9) and CATH (3), with a refinement resolution lower than 3Å to ensure high precision in structure determination. In the training sets of models from the first round, a length range of 51-76 amino acids has been chosen, as it is the Kunitz-domain length reported in the Prosite Prorule database (PRU00031 (6), the rule used to automatically annotate Kunitz domains on Uniprot); this length range assures also to retrieve polymer entities that have just one Kunitz domain, and do not have any other one, to avoid bad superimposition.

For the training set of model 1, contrarily to that of model 2, we have decided to exclude all the Kunitz domains that were described as variants, mutants and that contained mutations (artificially induced for experimental purposes). For the training sets of models 3 and 4 (round 2) we have modified the length range so that model 3 would have a stricter length range and model 4 a broader one, compared to that of model 1. Finally, in all the cases a grouping for sequence identity of 95% has been used to cluster the result and select representatives, avoiding redundancy: by grouping together sequences that have sequence identity lower than 100%, we put in the same cluster sequences that might be identical if not for few residues, hence we avoid a first level of redundancy, which would otherwise introduce biases in the models.

In each case, PDB has allowed us to retrieve tabular reports of our search results, containing the Entry ID, Auth Asym ID (the ID of the chain), resolution and annotation code (Uniprot associated ID). After a series of trimming procedures, for every training set we have ended up with a text file containing the PDB IDs (entry + chain containing Kunitz domain) that can be used to perform the multiple sequence alignment.

MSA and profile-HMM generation

The generation of the profile-HMM requires a multiple sequence alignment between the structures of interest; hence our second step has been the latter. Our MSAs have been performed with the online software PDBe Fold (5). Two multiple sequence alignments have been performed for each of our 4 models: the first MSA has allowed us to notice and remove structures that were badly superimposed and had a RMSD score greater than 2Å; the second MSA has allowed us to retrieve the best alignment. After retrieving them, the MSAs have been trimmed at the terminals (where we can find unaligned and gap regions) and some sequences have been removed: entries that were much shorter than the others, as well as entries that, after the trimming, contained identical sequences, to avoid a second level of redundancy. Finally, we have used the hmmbuild program of the HMMER package (4) with default options to construct the profile-HMM from the FASTA format of our cleaned MSA file.

Building the Test sets

During step 3, for each model a positive and negative testing set have been created: the positive set includes all the entries that are marked in the UniProt/SwissProt database as Kunitz domain, retrieved by entering as parameters of advanced search the Pfam identifier PF00014 and selecting those proteins that have been reviewed (which assures that we select the SwissProt database as search space). In each of the 4 model building procedures, the set of sequences that we have used for the training sets have been manually removed with computational procedures: the presence of these sequences, corresponding to the structures we have used in the training part, could have introduced redundancy, hence we have avoided a third level of potential bias. The negative set is composed by all the other proteins found in SwissProt, not marked as Kunitz-type proteins. Moreover, the training sets have been used as test sets to validate the correctness of the models. All the FASTA files were retrieved using the ID mapping tool of UniProt.

Model testing

For testing the models in step 4, we have made an extensive use of the hmmsearch program of HMMER, which aligns our HMM profile with amino acidic sequences in FASTA format retrieving the E-value of the alignment. Firstly, we have tested the model on our training set, a procedure that confirms the correctness of the model if the result of the E-value of the entries is high, meaning that our model can correctly recognize the sequences that have generated it (and that for sure are Kunitz-type domains). Next, we have united the negative and positive set and performed the hmmsearch step again, against this total set, retrieving the E-values. We have united the two test sets to avoid biases in the procedure, as the calculation of the E-value depends on the length of the considered database.

$$E - value = K \times m \times n \times E^{-\lambda \times S} \quad (1)$$

Where "K" is the dimension of the database, "m" the length of the model, "N" the length of the query, "λ" a parameter that represents the expected frequency of casual alignments, "S" the bit-score.

Thanks to computational procedures, we have been able to split again the negative and positive sets with the results of the alignments for each entry and have classified them, so that we can evaluate the performance of each model: entries of the positive set have been classified as 1 while the ones from the negative as 0. In the case of the negative set, most of the proteins have not been matched with the model, as they do not contain anything that resembles it. Hence, these entries have been automatically assigned with an E-value of 100.

The negative set and the positive set have been then further split into two smaller, randomized subsets, each containing half of the negative set and half of the positive, so that we can perform a cross validation test.

Performance evaluation

For the performance evaluation, we have written a proper python script that is able to compute a confusion matrix, thanks to the classifying procedure we have done beforehand: entries that obtained low E-values and have been marked as positive (1) are the true positives (TP), while entries that have a very high E-value and are labelled as negatives (0) are the true false (TN); false positives (FP) are be entries with low E-value labelled as negatives, while entries with high E-value labelled as positives are classified as false negatives (FN). The script has been used on each model for the 2 subsets and a third subset composed by the randomized union of the positive and negative tests from the previous section. Moreover, we have tested the models considering different thresholds of E-values. From the confusion matrix, the script can compute the Matthew's correlation coefficient (MCC), the accuracy score (ACC)and the F1 score.

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

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)})} \tag{3}$$

$$F1 = \frac{(Recall \times Precision \times 2)}{(Recall + Precision)} \tag{4}$$

The cross validation procedure implies the following steps: assessing the best approximated threshold (lowest E-value that has the highest MCC value) of the model on subset 1, the best approximated threshold on set 2, calculating the average of the two thresholds and testing the model on the total set at the averaged threshold. Once the best threshold has been found, the F1 score has also been calculated. By the comparison of MCC scores at the lower possible threshold and the computation of F1 scores, the best HMM profile between model 1 and model 2 was retrieved; by the end of round 2, the results coming from model 3, model 4 and model 1 were analyzed and the best HMM profile identified. ROC curves were also plotted for all the models.

Table 2. Representatives

model 1	1AAP:A 1BUN:B 1DTX:A 1KTH:A 1ZR0:B 3BYB:A 4NTW:B 4U32:X 5M4V:A 5PTI:A 1YC0:I 3M7Q:B 4DTG:K 4ISO:B 4U30:X 5YV7:A 6Q61:A 6YHY:A 1D0D:A 2UUY:B 2W8X:A
model 2	1AAP:A 1BUN:B 1DTX:A 1KTH:A 1ZR0:B 3BYB:B 4NTW:B 4U32:X 5M4V:A 1FAK:I 1G6X:A 1T7C:B 1YC0:I 2ZJX:B 3M7Q:B 3WNY:A 4DTG:K 4ISO:B 4U30:X 5NX1:D 5YV7:A 6Q61:A 6YHY:A 1D0D:A 2UUY:B 2W8X:A 3AUB:B 3AUD:A 3AUE:C 3AUG:A 3AUH:A 3AUI:A
model 3	1AAP:A 1DTX:A 1KTH:A 1TFX:C 1Y62:A 3BYB:A 4NTW:B 4U32:X 5M4V:A 5PTI:A 3T62:D 4ISO:B 4U30:X 5YV7:A 6YHY:A 1D0D:A 2UUY:B
model 4	1AAP:A 1BUN:B 1DTX:A 1KTH:A 1ZR0:B 3BYB:A 4NTW:B 4U32:X 5M4V:A 5PTI:A 1YC0:I 3M7Q:B 4BQD:A 4DTG:K 4ISO:B 4U30:X 5YV7:A 6Q61:A 6YHY:A 1D0D:A 2UUY:B 2W8X:A

Representatives: hereby are reported all the entries retrieved from the PDB advanced search. The coloured entries are the one that were removed during the MSA procedures.

Table 3. Removed sequences

model 1	1D0D 2UUY 2W8X 1YC0
model 2	1D0D 2UUY 2W8X 1YC0 3AUB 3AUE 3AUG 3AUH 3AUI 5NX1
model 3	1D0D 2UUY 2W8X 4NTW
model 4	1D0D 2UUY 2W8X 1YC0

Removed sequences: hereby are reported all the entries that have been removed from the MSA before the model building procedure. Coloured entries have been removed during the trimming procedure.

Results

Training sets

Our first round has aimed to find out if a model created without considering mutated structures could be better at identifying Kunitz-like domains than one trained also on variants. The initial training set for model 1 is composed of 21 representatives (71 Polymer Entities), while the one of model 2 by 32 (162 Polymer Entities). In the second round, we have aimed to modify the length range of model 1 to find the best model at different lengths of considered domains. Hence, training set 3 has 17 representatives (62 Polymer Entities) and set 4 has 22 (72 Polymer Entities). The PDB code of the polymer entities can be found in table 2.

MSA and profile HMM

During the multiple sequence alignment procedure, we have removed from the training sets proteins that consistently obtained bad RMSD values to retrieve better structural alignments. The removed sequences are listed in table 3.

It is possible to notice that entries 2UUY, 2W8X and 1D0D have been eliminated in most of the models. For model 3, we have decided to also remove entry 4NTW before running a second MSA. In each of the 4 trials, we have managed to align 3 elements of secondary structure (the two beta sheets and one alpha helix). Detailed results from the MSA can be found in the supplementary

Table 4. MSAs overview

	model 1	model 2	model 3	model 4
Overall RMSD	1.021	0.7443	0.9322	0.9568
Q-score	0.654	0.5169	0.7611	0.5387
aligned residues	51	39	52	50

Overview of the MSAs: the results of the 4 final multiple sequence alignments are reported in this table.

Table 5. Length of the tests set

	model 1	model 2	model 3	model 4
negative set	374	374	377	374
psitive set	569126	569126	569126	569126

Tests sets length: the number of entries of each testing set is reported in this table.

material. After the retrieval of the MSAs, the trimming procedure, in the case of model 2, has led to the removal of entry 5NX1, as it is much shorter than the others and could have given problems in the model generation procedure. Finally, we have had to remove sequences that after the trimming procedure were found to be identical: for model 1,2 and 4 we have removed the entry 1YC0, identical to 4ISO but with lower resolution, from model 2 we have also removed 3AUG, 3AUH, 3AUI, 3AUE, all identical to 3AUD. For this procedure we have used the help the Uniprot tool Align, which alerts us if two sequences are identical.

In table 4 are reported the overall RMSD scores of the alignments and number of aligned amino acids. The profile HMMs can be found in the supplementary material.

Testing sets

Looking at the testing sets, for each model the negative set has the same length (being in every case the full set of SwissProt entries excluding those annotated as Kunitz domains), while the positive set has a variable number of entries: this is because every time we have had to remove the list of redundant entries, meaning the entries that were used to train the models, to avoid biases. In table 5 are reported the lists of redundant entries from every model.

The first testing procedure, the evaluation of the model against the list of redundant proteins (conformed by the UniProt IDs corresponding to the training set), have always given positive results. Hence, the models have been able to identify correctly the Kunitz domains used for the training procedure. The results of the testing procedure against the union of negative and positive sets are shown in the next paragraph.

Evaluation of the performance

The performances of model 1 (no mutants), model 2 (with mutants), model 3 (no mutants low range), model 4 (no

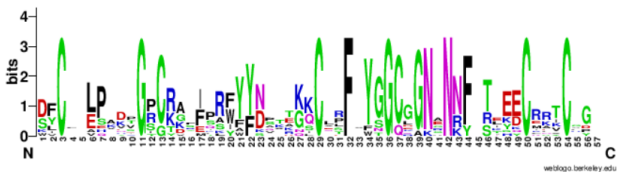

Fig. 4.: Sequence logo of model 3 (no mutants, low range)(1).

Table 6. Best performances at average thresholds

	model 1	model 2	**model 3**	model 4
TH	1E-04	1E-03	1E-04	1E-03
ACC	0.99999	0.99999	0.99999	0.99999
MCC	0.99732	0.99732	0.99734	0.99732
F1	0.997319	0.997319	0.99734	0.997319

Best performances: results of the 4 model at their respective best threshold. Higlightened the best models.

Table 7. Best performances of model 1 and 3

	model 1	model 3
TH	6E-05	7E-05
ACC	0.99999	0.99999
MCC	0.99732	0.99734
F1	0.997319	0.99734

Best performances of model 1 and 3 according to their best thresholds. Best model highlightened.

mutants high range) at different thresholds can be found in the supplementary material, while here in table 6 are reported the best performances. The approximated best threshold has been retrieved using the cross validation procedure.

Discussion

The idea for this project was to test, first, if the presence/absence of mutant Kunitz-type domains could affect the performance of the model. Indeed, it has been shown that a broader variety of structures, given by the presence of mutations in the Kunitz domains, did not lead to a better model. This might indicate that mutated structures, considered in model 2, introduce too much variability in the model, therefore lowering its performance. Moreover, by modifying the length range of the considered domains, this study has shown that a smaller range leads to a better classifier for higher thresholds, while a broader range leads to a better classifier for lower thresholds, but with lower performances. The model obtained with a middle length range, model 1, seems to be better than model 3 in a small E-value threshold interval, between 7E-05 and 6E-05. These results may be explained by the fact that, by selecting a lower length range, we select fewer, more similar structures, hence we obtain a model with a better resolution. In spite of this, the model doesn't perform on lower E-value thresholds as good as the models which are trained on a broader number of sequences. We also have to consider the fact that, by trimming the MSA, in the end the models have more or less the same length, hence lower/broader length range has an impact only on the structure retrieval.

For what concerns the multiple sequence alignment procedure, three sequences (1D0D:A, 2UUY:B, 2W8X:A) consistently obtained a RMSD score greater than 2, therefore we had to remove them from every dataset. Moreover, we eliminated entry 3AUB:B from model 2 and entry 4NTW:B from model 3 for the same reasons. By looking carefully at the first 3 entries, and double checking on Uniprot database, we have noticed that they all belong to a tick species; this is indeed an interesting occurrence, as it is known from previous studies that Kunitz domains from anticoagulant protein of these insects have a different protease inhibition mechanism, hence a folding variation, even if they

Table 8. Performances at best overall threshold

	model 1	model 2	model 3	model 4
TH	6E-05	6E-05	6E-05	6E-05
ACC	0.99999	0.99999	0.99999	0.99999
MCC	0.99732	0.99462	0.99600	0.99463
F1	0.99733	0.99462	0.99600	0.99462

Performances of the 4 models at the overall best threshold (the threshold that obtained the best overall values for at least 1 model). Best model highlightened.

Table 9. Performances at lowest acceptable threshold

	model 1	model 2	model 3	**model 4**
TH	1E-07	1E-07	1E-07	1E-07
ACC	0.99998	0.99998	0.99999	0.99999
MCC	0.99194	0.99194	0.99334	0.99463
F1	0.99191	0.99191	0.99332	0.99462

The table shows the performances of each model at the lowest acceptable threshold, the threshold of E-value under witch the models started accumulating too many False negatives to be considered accurate. Best model highlightened.

maintain the 6 Cysteines (17). On another note, the 3AUB entry is a variant of the Kunitz domain, hence the mutations (A14G and A38V) probably led to a change in the folding pattern. Finally, for the 4NTW:B entry, we have to specify that the RMSD was below the acceptable threshold, hence the entry was suitable for our experiment; we decided to remove it because the RMSD was higher than all the other entries, and this had a confirmation in the MSA itself, as the presence of the entry introduced in the alignment two gaps.

Looking at the testing sets, at the positive set, which contains the UniProt annotated Kunitz domain, it is noteworthy to mention that after removing the entries that were used for the training sets, the number of entries was always 374 (except for model 3, 377), meaning that 16 entries were removed (the number of Pfam annotated Kunitz domain on UniProt is 390). This is because some PDB entries do not have any UniProt associated ids, also some of them are linked to the same UniProt entry.

Finally, looking at the performance evaluation, we can see that in all our 4 models, five specific entries were always classified as false negatives, therefore scored low in the model testing, lowering our overall performance. False positives, on the other hand, were not identified until a higher threshold. Looking at two entries in particular, O62247 and D3GGZ8, are remarkable: by looking at their UniProt page we can see that they both belong to the bli-5 gene and it appears that they do not have a serine protease inhibition activity, or at least the scientists are not sure if their activity is proper, as they lack catalytic sites (18). It is possible that these proteins were wrongly annotated or that they represent distantly related homologues, as by looking directly at the alignment between the models (we examined the alignment with model 3, supplementary material) and the sequences, they lack two of the 6 Cysteines, which may explain these results.

The same cannot be said for the other 3 false negative entries, Q11101, D9IFL3, and P86963: looking at them carefully, it is possible to notice that their annotation score is very low (2/5) and that their serine protease inhibition activity is automatically annotated and not manually reviewed. On top of that, by blasting the sequences against the UniProt database, the results show that they have no close similar entities in term of sequence identity and resulting entries all belong to UniProt TrEMBL database. Looking at the alignment of the models against the sequences, it is possible to see that they have the 6 conserved Cysteines. This means that we have no reason to believe these entries, although having a low annotation score, are not Kunitz like domains, maybe from distant homologues, and removing them from the database, to improve our results, has no sense. Indeed, we have performed another testing procedure on our models with the testing dataset deprived of entries O62247 and D3GGZ8. The results are roughly the same as with the original set, and the removal of this uncertain Kunitz domains leads to a slight improvement of the overall MCC score and F1 score, but doesn't change the approximated best thresholds (view supplementary material).

Conclusion

In conclusion, our aim to find a suitable profile-HMM to annotate Kunitz type domains has been successfully achieved. The results indicate that the models obtained from training sets without mutant Kunitz domains are better at classifying new sequences. Moreover, the model built considering a small length range for the training structures (50-60 AA) obtained the overall better performance at high E-value thresholds, while the model obtained using a broader range (50-80 AA) is a better classifier at lower E-value thresholds, but with lower overall performances. As a side note, some entries consistently scored as false negatives in all of our models, which might be due to annotation errors; on the contrary, it appears that, as our models didn't find false positives, currently there aren't Kunitz domains that are not annotated as such from Pfam in UniProt/SwissProt.

References

1. Weblogo. URL: `https://weblogo.berkeley.edu/logo.cgi`.
2. Merops, 2017. URL: `https://www.ebi.ac.uk/merops/`.
3. Cath, pancreatic trypsin inhibitor kunitz domain, 2019. URL: `http://www.cathdb.info/version/latest/superfamily/4.10.410.10`.
4. Hmmer, 2019. URL: `http://hmmer.org/`.
5. Pdbe fold, 2019. URL: `https://www.ebi.ac.uk/msd-srv/ssm/`.
6. Prorule, pru00031, 2019. URL: `https://prosite.expasy.org/rule/PRU00031`.
7. 1bpt pdb, 2023. URL: `https://www.rcsb.org/structure/1BTP`.
8. Interpro, pancreatic trypsin inhibitor kunitz domain, 2023. URL: `https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002223/`.
9. Pfam, bovine pancreatic trypsin inhibitor domain, 2023. URL: `https://www.ebi.ac.uk/interpro/entry/pfam/PF00014/`.
10. Quickgo, go:0004867, 2023. URL: `https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0004867`.
11. Rcsb pdb, 2023. URL: `https://www.rcsb.org/`.
12. Uniprot, 2023. URL: `https://www.uniprot.org/`.
13. Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

14. Juliana Cotabarren, Daniela Lufrano, Mónica Graciela Parisi, and Walter David Obregón. Biotechnological, biomedical, and agronomical applications of plant protease inhibitors with high stability: A systematic review. *Plant Science*, 292:110398, 2020.
15. Kazuho Ikeo, Kei Takahashi, and Takashi Gojobori. Evolutionary origin of a kunitz-type trypsin inhibitor domain inserted in the amyloid β precursor protein of alzheimer's disease. *Journal of molecular evolution*, 34:536–543, 1992.
16. Manasi Mishra. Evolutionary aspects of the structural convergence and functional diversification of kunitz-domain inhibitors. *Journal of Molecular Evolution*, 88(7):537–548, 2020.
17. Guido C Paesen, Christian Siebold, Karl Harlos, Mick F Peacey, Patricia A Nuttall, and David I Stuart. A tick protein with a modified kunitz fold inhibits human tryptase. *Journal of molecular biology*, 368(4):1172–1186, 2007.
18. Antony P Page, Gillian McCormack, and Andrew J Birnie. Biosynthesis and enzymology of the caenorhabditis elegans cuticle: identification and characterization of a novel serine protease inhibitor. *International journal for parasitology*, 36(6):681–689, 2006.
19. Neil D Rawlings, Dominic P Tolle, and Alan J Barrett. Evolutionary families of peptidase inhibitors. *Biochemical Journal*, 378(3):705–716, 2004.
20. SCOP. Scop, small kunitz-type inhibitors bpti-like toxins, 2021. URL: `https://scop.mrc-lmb.cam.ac.uk/term/4003337`.
21. Hiroshi Shigetomi, Akira Onogi, Hirotaka Kajiwara, Shozo Yoshida, Naoto Furukawa, Shoji Haruta, Yasuhito Tanase, Seiji Kanayama, Taketoshi Noguchi, Yoshihiko Yamada, et al. Anti-inflammatory actions of serine protease inhibitors containing the kunitz domain. *Inflammation research*, 59:679–687, 2010.
22. Hyun Kyu Song and Se Won Suh. Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from erythrina caffra and tissue-type plasminogen activator. *Journal of molecular biology*, 275(2):347–363, 1998.
23. Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009.