Project report

HMM profile for the Kunitz domain based on a structural alignment

Krsto Vujovic, MSc Bioinformatics at UNIBO, Bologna University.

Abstract:

Kunitz motif is a cysteine-rich peptide chain of~60 amino acid residues with alpha and beta fold, stabilized by three conserved disulfde bridges. An extensive dataset of amino acid variations is found on sequence analysis of various Kunitz peptides. Kunitz peptides show diverse biological activities like inhibition of proteases of other classes and/or adopting a new function of blocking or modulating the ion channels. Based on the amino acid residues at the functional site of various Kunitz-type inhibitors, it is inferred that this ‘fexibility within the structural rigidity’ is responsible for multiple biological activities.

We developed a HMM model based on structural alignment for predicting Kunitz domain type BTPI (PF00014 Pfam Identifier).We showed that this is good way for building model for Kunitz domain, which give high accuracy and Matthews correlation coefficient of 0.99% (almost 100%)

Introduction

Kunitz type proteins are an important group of ubiquitous protease inhibitors found spanning the evolutionary tree from microbes to mammals. These proteins can have single or multiple Kunitz inhibitory domains linked together or associated with other domain types.[i] Kunitz-type serine protease inhibitors are found in several organisms including animals, plants, and microbes. Kunitz-domain inhibitors known from animal sources are classified under the inhibitor family I2, Clan IB according to the MEROPS database.[ii]

Bovine pancreatic trypsin inhibitor (BPTI) is the classic member of this family of proteins and was the first Kunitz-type protease inhibitor described (Kunitz and Northrop, 1936). BPTI has relatively broad specificity inhibiting trypsin as well as chymotrypsin and elastase-like serine (pro) enzymes. Kunitz inhibitors are thus known as BPTI-like proteins and belong to the I2 family of peptidase inhibitors. Kunitz inhibitors may contain a single domain (e.g. BPTI) or the domain may be repeated twice (e.g. *Bikunin*), three times (e.g. tissue factor pathway inhibitor-TFPI) or even more (there are 12 domains in the *Ancylostoma caninum* (hookworm) Kunitz protease inhibitor) to form a multi-domain, single-chain inhibitor able to interact independently with several protease molecules at their reactive sites belonging to separate domains.[i]

They 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 (Fig 1 (a)). The disulfide bridges maintain the structural integrity of the inhibitor and also present the protease-binding loop at its surface. A highly exposed P1 active site residue at position 15 is usually arginine (Arg) or lysine (Lys) inserts into the S1 site of the cognate protease and is the primary determinant of the specificity of serine protease inhibition.[ii]

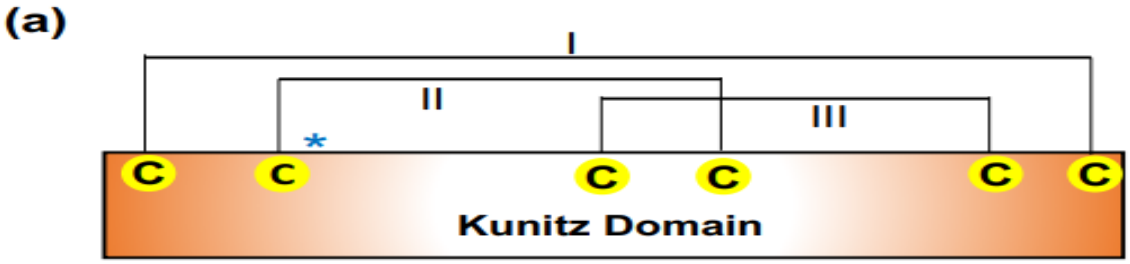

Fig 1. (a) Structure and activity of Kunitz-domain inhibitors (a) Schematic representation of Kunitz-domain inhibitor showing the disulfide bonding pattern of 1–6, 2–4, 3–5. Conserved cysteine residues are marked with yellow and the P1 active site residue, the primary determinant of the specificity of serine protease inhibition is marked with a blue asterisk.

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 (viz. trypsin/chymotrypsin/elastase/kal-

likrein).[ii] In addition to their major action in serine protease inhibition, some Kunitz proteases can act as ion channel blockers and are known as Kunitz-type toxins (KTT). They are frequent components of the venoms from poisonous animals including snakes, sea anemones (e.g. *Anthopleura elegantissima*, cone snails (e.g. *Conus striatus*), tarantulas (*Ornithoctonus spp.*), scorpions (*Lychas mucronatus*) and the cattle tick *Boophilus microplus*.[i]

Kunitz-domain sequence is found in the precursor amyloid β-protein (APPI) which accumulates in the neurotic plaques and cerebrovascular deposits of patients with Alzheimer's disease, the most common neurodegenerative disorder (Fig 2.).[ii]

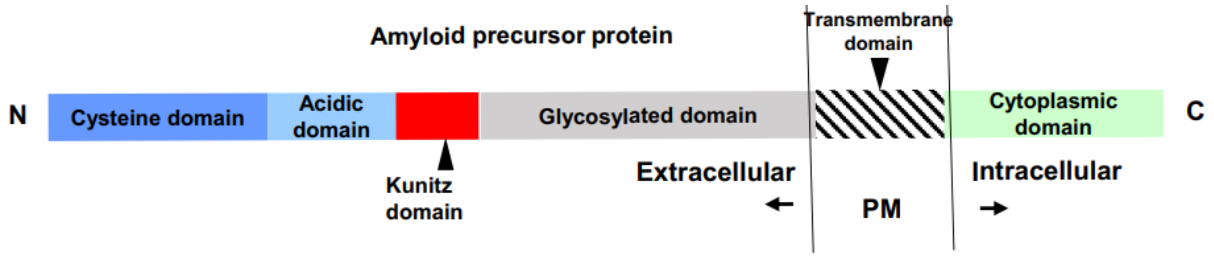

Figure 2. Amyloid β-protein precursor (100–135 kDa) is an integral membrane glycoprotein that contains the Kunitz domain as an insertion between the acidic domain and the glycosylated domain. Alternative splicing results in overproduction of Kunitz-inhibitor-containing species in Alzheimer's patients.

Kunitz-type domains exist in multiple forms in numerous tissues imparting proteins with specific (serine) protease inhibitory function: the Kunitz-type toxin in venomous animals like snakes, spiders, and scorpions, mammalian inter-alpha-trypsin inhibitors, domain found in Alzheimer's amyloid β-protein in humans, domains at the C-termini of the alpha-1 and alpha-3 chains of type VI and type VII collagen and tissue factor pathway inhibitor (TFPI) (Fig. 1 (d)).[ii]

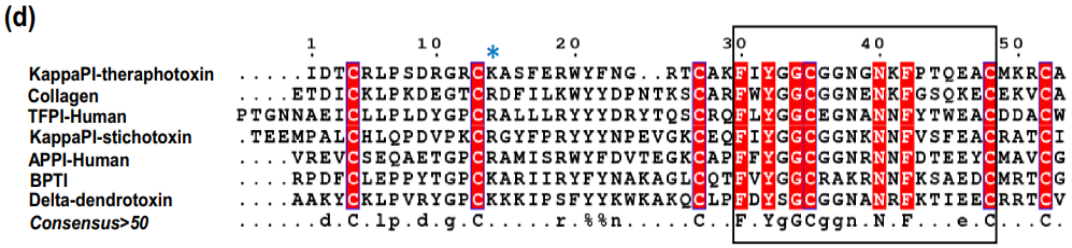

Fig 1 (d): d Amino acid sequence alignment of Kunitz-domain inhibitors from different source organisms: BPTI from *Bos taurus* (P00974); Delta-dendrotoxin from *Dendroaspis angusticeps* (P00982); KappaPI-theraphotoxin from *Haplopelma schmidti* (P68425); KappaPI-stichotoxin from *Stichodactyla haddoni* (B1B5I8); Kunitz domain from human (*Homo sapiens*) amyloid β-protein precursor (P05067); Kunitz domain from human (*Homo sapiens*) collagen alpha-3(VI) chain (P12111); Kunitz domain from human (*Homo sapiens*) Tissue factor pathway inhibitor 2 (P48307). 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 (Color figure online)

Methods and Materials:

Dataset Generation

First step into building HMM model is to gather the dataset from which we will train our model. The dataset of protein structures was obtained by an advanced search option through PDB[iii] database and this file can be found in the supplementary material under name (rcsb_pdb_custom_report_20230421014634.csv). The following constraints were selected: identifier of the PFAM[iv] Protein family (PF00014 is identifier for Kunitz_BPTI type), sequence length in range 49-80 (upper included), resolution of 3D structure smaller or equal to the 3 Å, polymer entities grouped by sequence identity of 95% and displayed as representatives. The identifier serves as a filter for those sequences that are containing Kunitz domain, sequence length serves as an filter for selecting only structures with single Kunitz domain (as above described typically length of Kunitz domain is 50-70 amino acid residues), because as above mentioned we can have multiple Kunitz domain in a single sequence which can be sequentially positioned or on different subunits which would be problematic for the model analysis. Resolution constrains serves as filter for only protein structures that are defined within "good" resolution and at the end we grouped by sequence identity of 95% because we wanted to exclude redundancy of having multiple PDB structures of the same protein.

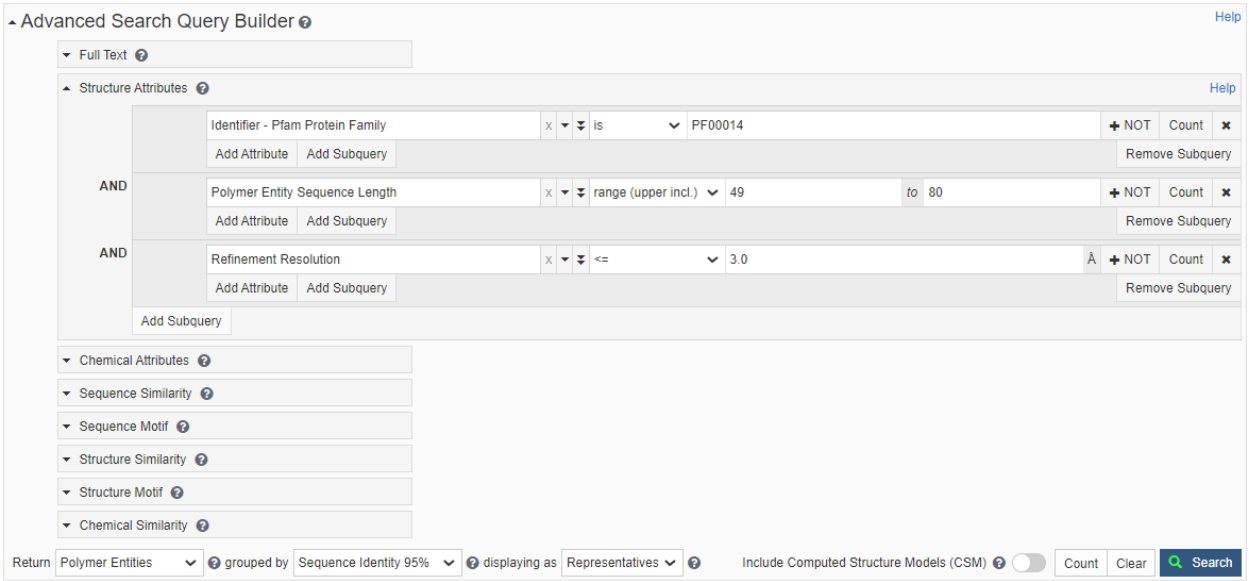

Fig 3. Screenshot of the parameters(constrains) for advanced search on the PDB Database.

Structural alignment with PDBeFOLD

Protein structures dataset was cleaned and saved as list_pdbefold.txt which can be found in the supplementary material, file was uploaded on web application PDBeFOLD for multiple structural alignment with default parameters. The output alignment was downloaded and can be found in supplementary material as aln_3d.fasta.seq file.

Building the HMM Model from structural alignment with HMMER

HMM Model was build with hmmbuild function from HMMER version 3.3.2 Nov 2020.

More about hmmbuild function and hmmsearch function which will be mention later, can be found in supplementary material in document HMMER User Guide's, the output of the hmmbuild can also be found as ali_3d.hmm file in the supplementary material.

Generation of the testing dataset

For negative test dataset, was generated with the advanced search on Uniport[v], on 21 April, version release-2023_01, the parameters were (**NOT**) PF00014 (Pfam Domain) and reviewed (**YES**), this negative testing dataset can be found at supplementary material as nonkunitz.fasta file.

Positive test dataset was generated by combining two files human_kunitz.fasta which contains 18 sequnces and kunitz_nonhuman.fasta which contains 364 sequences, both of the files were generated in similar project for blast exercise, the final file can be found in supplementary material under name kunitz.fasta which contains 384 sequences (366+18), we could also generate this file by advanced search on Uniport with parameters PF00014 (Pfam Domain) (**YES**) and reviewed (**YES**), although bear in mind that there is new version of **Uniprot release-2023_02** so search could contain more sequences then ours.

Before we go further, we needed to do subtraction of the 33 training sequences from this positive test dataset to make sure that we don't have bias in our testing set. First we used ID Mapping tool on Uniport to map these 33 PDB sequences to the 48 Uniport/SwissPort codes (one PDB can match to multiple Uniport's, this file is called list_pdb2uniprot.txt in the supplementary material), because our kunitz.fasta contains Uniport codes. Then we used comm Linux command which creates output like Ven Diagram, in the first column gives A minus B , second column B minus A and third column contains intersection between A and B (16 sequences). Finally, positive dataset can be found in the supplementary material under name kunitz_clean.txt which contains 368 structures (384-16). These 368 protein codes we mapped from Uniprot AC-ID to the Uniprot/SwissProt codes and saved that file as final positive test dataset kunitz_clean.fasta file.

Optimization with two k-fold cross validation

On both files from previous step (kunitz_clean.fasta and nonkunitz.fasta) hmmsearch function with options (--max),(--noali) was run, and we got corresponding hmm.search files which we cleaned and added classifiers, final files can be found in the supplementary material under name nonkunitz_original.class (for negative class) and kunitz.clean.class (for positive class).

Then we randomized both classes and put half of positives and negatives in set_1_original.txt file, the other half of positives and negatives was put in set_2_original.txt file. The python script Optimization.py was written, that upon call, with given set and threshold value, calculates confusion matrix, accuracy and MCC (Matthews correlation coefficient[vi]). Two k-fold cross validation[vii] was then run and best threshold was selected which will be described under results and discussion.

Results and Discussion:

First results that needs to be described are coming from second step in the workflow, in structural alignment, the average RMSD value of structural alignment of 33 protein sequences was 0.7349 which is very good value because its under 1 Å, meaning that alignment is of good quality so we can expect that building HMM model from this structural alignment will result in good predictions. On another note, we used WebLogo to represent sequence profile (Fig 3.) out of structural alignment file.

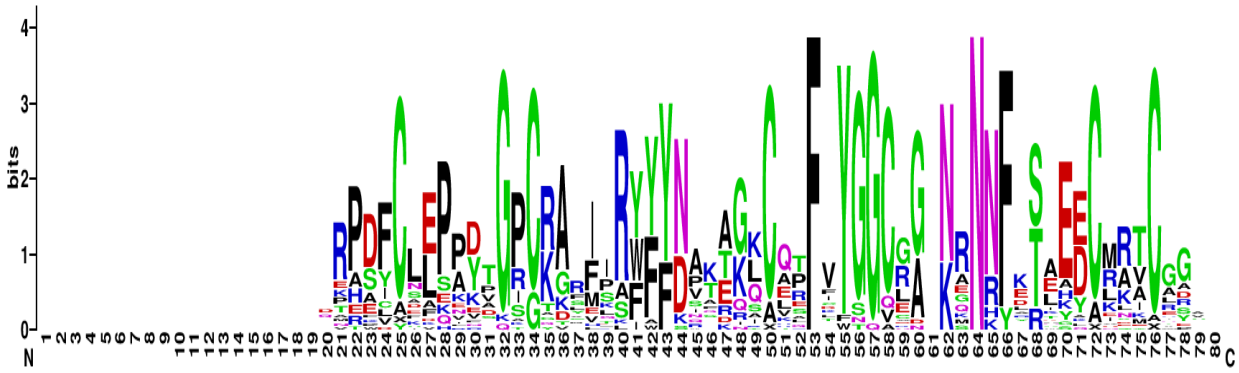

Fig 3. Schematic view of sequence profile that was generated out of structural alignment file.

From the sequence profile we can see which residues are really conserved and ones out of more importance here are the famous six cistein's that are very characteristic for the Kunitz Domain but if we compare this sequence profile to the one alignment shown on the Fig 1 (d) (BPTI type) we can see that we also have conserved characteristic Phenylalanine, Tryptophan and Asparagine residues along the sequences.

From generating HMM model with hmmbuild function, we can say that input file contained 80 aligned columns (alen) from 33 protein sequences. HMMER turned it into 58 consensus positions which would mean (80-58=22) gap-containing alignment columns to be insertions relative to consensus.

To summarize generation of test dataset process, negative test dataset contains 568.829 proteins that were extracted from Uniprot database, for the positive test dataset, we had 384 positive examples from which 16 where excluded (matched with training set) so at final we have 368 proteins.

Last part of workflow was optimization with two k fold cross validation, where first we run hmmsearch function (which makes a search through the given database (in this case test sets)), as output of this function, the most important part is the second section where is the sequence top hits list. It is a list of ranked top hits (sorted by E-value, most significant hit first), formatted in a BLAST-like style.

Here important to mention is that given the positive test dataset we got 368 hits (like expected) and highest e value was 0.0027 (2,7e-03) which is good cut off value for the positive set and of course all of these examples were set with classifier 1 (binary classification where 1 means contains Kunitz domain and 0 does not contain Kunitz domain).

For negative test dataset we got 39 hits out of 568.829 proteins, between those 39 hits the lowest e value is 4,7e -24 and highest e value is 27. After cleaning the output search file, we needed to add to that file (nonkunitz_original.class) remaining 568.790 that weren't matched, because these are also negatives just the e value is too high to pass given threshold on the hmm search function. We added a random e value of 100 on all examples of 568.790 and then concatenated two files and added classifier 0.

For optimization part, we created two sets both containing half of positive and negative test class, its important to mention that these two sets are not identical, due to uneven number of negative test class (568,829) in set_2_orignial.txt file we have one more example then in set_1_original.txt file. Python script was created for purpose of calculating confusion matrix. Confusion matrix represents matrix (2x2), where on the diagonal of the matrix we have **True Negative** and **True Positive** examples as shown on the picture below, on the reversed diagonal we have **False Positive** and **False Negative** examples. Confusion matrix is shown on the table below.

Table 1. Confusion Matrix

Predicted values	Actual values		
	Total examples	Negative	Positive
	Negative	**True Negative**	**False Negative**
	Positive	**False Positive**	**True Positive**

From confusion matrix we are able to calculate ACC(Accuracy) which is not so highly informative, could be bias if you have unbalance dataset like in our case.

$$\textit{Accuracy (ACC):}$$

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

Fig 4. Formula of Accuracy

On other hand MCC (Matthews correlation coefficient) gives much more broader picture of classes, because its not affected by dataset unbalance, MCC is near 0 in case of random cases, it is 1 in case of perfect prediction and -1 in case of completely wrong prediction. In our optimization process we will determine best e-value threshold based on MCC value as determining factor how well our model predicts.

Matthews Correlation Coefficient (MCC):

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

Fig 5. Formula of MCC

Optimization process was organized as follows : for set_1_original.txt we tested 20 e values in range of 10^{-1} to 10^{-20} then best MCC value determined optimized threshold for set 1 which was then tested on set 2. The same was done for the set_2_original.txt and at end average value is taken as the optimal one in which we combined both sets and tested the given threshold.

Table 2. Optimization process

	SET	THRESHOLD	ACC	MCC
Optimization	set1	1e-03	0.99998594	0.98929754
TEST	set2	1e-03	0.99998945	0.99186829
Optimization	set2	1e-02	0.99999297	0.99460562
TEST	set1	1e-02	0.99998243	0.98667515
Final	set1+set2	1e-02	0.99998770	0.99061660

As we can see best e value for set 1 was 1e-03 and 1e-04 while for set 2 best value was 1e-02. For set 1 we had choice and I chose 1e-03 as that value is closer to the optimum of the set2 and gives better results overall. As average value of the optimal value between 1e-03 (set_1) and 1e-02 (set_2) is 1.5e-02 but this does not give any significant difference from 1e-02 (rounded number).

Now if we look at ACC and MCC we can se that model is pretty good at predicting, both ACC and MCC are 0,99%. But if we want to have closer look of what makes this 1% problematic, we need to gather information about confusion matrix which is shown below.

Table 3. Confusion matrix

Predicted values	Actual values		
	Total examples	Negative	Positive
	Negative	568 822	0
	Positive	7	368

From confusion matrix we can deduce following things, model have problems with positive examples, to be exact we have 7 examples that are characterized as false positive, while negative group is perfectly predicted. These seven example if we look at column of the actual values we can see they can be found in our negative set class (nonkunitz_original.class). If we sort this file by reverse values we can find the 7 examples with corresponding e value, remember that we said lowest e value in negative test dataset was 4,7e-24.

List of lowest 7 values in nonkunitz_original.class file is presented in the table 4:

Table 4. List of the 7 examples that are wrongly predicted.

Uniprot Code:	E value:	Classifier:
C0HLA7	4.7e-24	0
C0HLA5	5e-24	0
C0HLA6	1.1e-23	0
C0HLA8	2.3e-23	0
C0HLA9	1.1e-22	0
C0HLB0	2.5e-14	0
P84555	0.0026 (2.6e-03)	0

We can see that C0HLA is same for 5 examples, which already tells us that they are very similar proteins, also we can see that these proteins have highest e values in the negative class. From Uniprot's page we can see that all 5 examples are proteins in snakes, which was mentioned in introduction part that there are Kunitz domain in some venomous snakes. This is clear example, that we can confirm if we check their uniport page under section family and domains as shown on Fig 6.

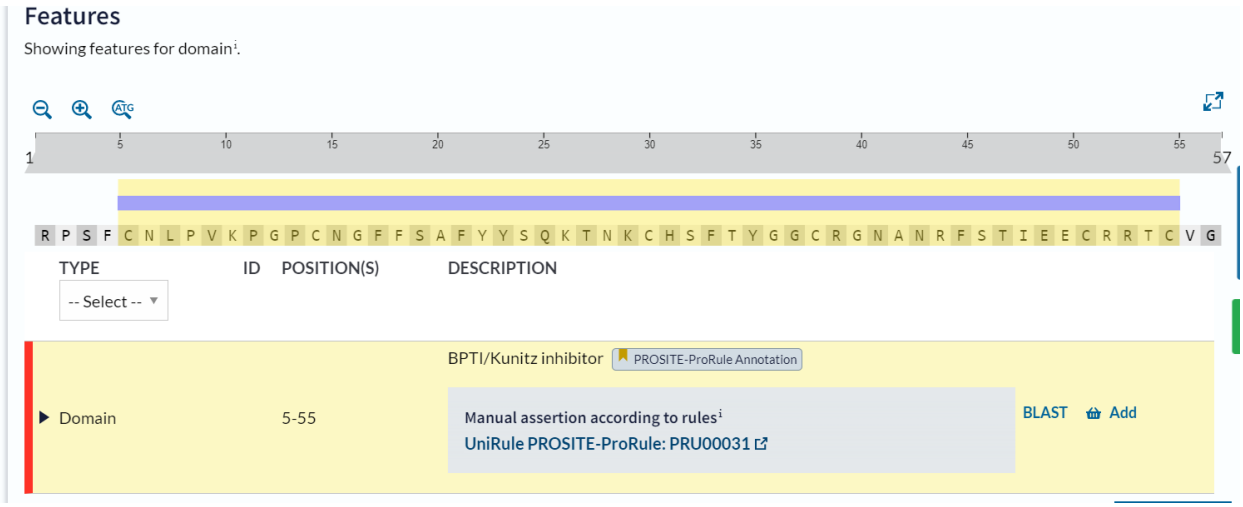

Figure 6. Uniprot page of C0HL5, section family and domains

We can see that evidence for the BPTI Kunitz domain is manual assertion according to rules, which means it was reviewed. Now as for the reason why this example is shown in our negative set if its manually asserted?! The reason could be that manual assertion is only done recently with current

version of Uniprot release 2023_02. So to confirm this hypothesis we downloaded new negative set under same constrains and make a new nonkunitz.class to check if the following examples are still there. This file nonkunitz.class had 569 126 (more then previous one with 568k) out of these only 33 examples got match after hmmsearch function (which means six less examples) and if we check the lowest e values in this negative class we found lowest value of example P84555 2.6e-03. This is the last example from our table 4, and we can clearly deduce that all six examples of Kunitz BPTI domains that are found in venomous snakes, got manual assertion and were marked as proteins that are containing PF00014 Pfam domain.

As for the P84555 protein, it is clear that it is Kunitz-type serine protease inhibitor, because functionality evidence confirms it and also we have manual assertion according to prosite rules in the uniport page under section family and domains but if we go InterPro [viii] website which contains Pfam domain database we can gather more information and insights.

Fig 7. InterPro page of Uniprot P84555 protein

From Fig 7 we can clearly see that we have a matching of Kunitz_BPTI domain, but the length of the segment is from 4-27, which tells us that it is not a normal and fully characteristic Kunitz Domain (they are often from 50-70 amino acid residues long), but fraction of the sequence. That's why e value is not too strong like in other examples like proteins from snakes, but again it is enough to get recognized with decent e value from our HMM model.

Conclusion:

To summarize everything, we saw that with dataset of good PDB structures, we can make high quality structural alignment with PDBeFold (Avg. RMSD is under 1 Å) and that from visual representation of sequence profile have conserved all positions that are characteristic for Kunitz domain BPTI type . Upon building model with HMMER function hmmbuild from structural alignment, and testing it against positive and negative datasets. We saw from MCC values and confusion matrix that we can get very high prediction accuracy (almost 100%) . But we also

understood that we could improve model by downloading new negative set from recent release 2023_02 where all 6 examples of snakes proteins, which were very problematic as **False Positive** examples, got manually annotated and would no longer pose problems. In confusion matrix only problem, were False positive examples but now by eliminating 6 out of 7 examples with new dataset, maybe with new optimization we could pick a threshold that could give MCC value of 1 with 100% accuracy. This could be a potential new way in which model should further be developed.

References:

[i] Ranasinghe and McManus, "Structure and Function of Invertebrate Kunitz Serine Protease Inhibitors."

[ii] Mishra, "Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors."

[iii] Burley et al., "Protein Data Bank (PDB)."

[iv] Bateman et al., "The Pfam Protein Families Database."

[v] Consortium, "UniProt."

[vi] Chicco and Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation."

[vii] Rodriguez, Perez, and Lozano, "Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation."

[viii] Hunter et al., "InterPro."