Automatic classification of kunitz domain using Hidden Markov Model

Lui Maria[1]

Abstract

Kunitz-type proteins like the bovine pancreatic trypsin inhibitors have the main function of breaking down of blood clots and they perform this function throught a very important domain called Kunitz domain. The Kunitz domain infact characterize a very numerous protein family of protease inhibitors of pharmacological interest since they are the main component of the Aprotinin drug, meant to reduce bleeding after surgery . The aim of this paper is to develop a method to annotate Kunitz domains in uncharacterized proteins.
In order to do it we started from available structural informations of this kunitz type protease inhibitors and then we build a profile Hidden Markov Model (HMM) for the Kunitz domain. The HMM profile is the tool of choice for protein family research, it is a probabilistic generative model that specifies position-specific letter emission distributions and also position-specific insertion and deletion probabilities to describe a sequence family. In order to represent the Kunitz domain we generated an HMM profile based on a structural alignment using HMMER. Finally we tested the performance of our model and we assessed it correctly discriminates the proteins belonging to the Kunitz family from the other ones that were classified as not belonging to the family. The overall performance of our method was computed at the threshold of 10^{-6} and it shows a good sensitivity (TPR =1.0) and accuracy (ACC=0.999). Moreover the computed Mattew Correlation Coefficient for our model is equal to 1 (MCC=0.997) and the number of false positive rate (FPR) was computed at zero; meaning that the overall performance of our method is good.

1 Introduction

A crucial step in blood coagulation is the conversion of prothrombin into its active form, thrombin; then the latter converts fibrinogen into fibrin and leads to the formation clots that impede blood loss. [3] The opposite process is the fibrinolysis since it prevents blood clots from growing and becoming problematic, breaking down the fibrin clot. Its main enzyme is plasmin, it is present in blood and it degrades many blood plasma proteins and cuts the fibrin mesh . [8] Since it is a central mechanism in cardiovascular disease, an unregulated proteolysis will have undesirable effects in thrombotic and atherosclerotic processes and in the response to ischaemia-reperfusion injury. Therefore protein that play an important role in preventing the fibrinolysis have been deeply studied and are called protease inhibitors.
One of the most numerous protease inhibitors family is the Kunitz-type, usually serine protease inhibitors.

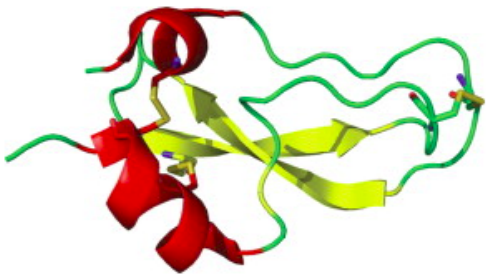

Figure 1: Structure of BPTI. In this structure (PDB ID 1BPI) α-helices are shown in red, β-sheets in yellow, and loop regions in green. Three conserved disulfide bonds are shown as sticks. [17]

Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI), whose kunitz domain sequence are compared in Figure 3. Proteins are generally composed of one or more functional regions, commonly termed domains. Kunitz domains are the active domains of those proteins that prevent access of the serine protease to its physiological substrate through an insertion of a residue into

the active site cleft(Arg-24 in BPTI kunitz domain,that is frequently used as a model of this family.) [9]
The pancreatic trypsin inhibitor has a low relative molecular mass, a basic isoelectric point and one or several inhibitory domains with a broad spectrum of activity toward serine proteases. [5]
The Kunitz family of serine protease inhibitors are relatively small, they consist of between 50 and 70 amino acids and a molecular weight of 6 kDa, they are α/β proteins since they adopt a conserved structural fold with two antiparallel β-sheets and one or two helical regions that are stabilized with three disulfide bridges(Figure 1).

The stabilizing disulfide bonds are reported in Figure 2 of BPTI (bovine pancreatic trypsin inhibitor) .
A very well solved kunitz domain that we used in our model is the one of the a3 chain of human type VI collagen (PDB: 1KTH) . It is a single amino-acid residue chain with three disulphide bridges, solved through X-Ray diffraction with a resolution of 0.9 Å. [4]

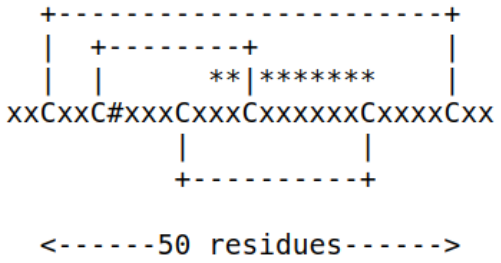

Figure 2: kunitz domain:conserved cysteine involved in a disulfide bond are remarked (C) togheter with active site residue (#) and the position of the pattern (∗).

2 Materials and methods

2.1 Datasets

In order better understand the function of kunitz type protein we first identified the domain that occurs within this family using Pfam. The Pfam database (release 32.0 ,September 2018) [13] was used to to derive the Pfam id for the Kunitz BPTI domain (Pfam:PF00014) since it is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Then we performed a search in the Protein Data Bank (PDB) to extract the structure of the protein containing this domain. [2, 6, 7] We searched on PDB for Pfam Accession Number PF00014, structures solved with X-Ray crystallography with a resolution between 0.0 and 3.0 and wild type protein; the resulting query structures were 173 (reported in Table S1 in Supplementary Materials).The resulting 173 structures have been sorted for their resolution, it allowed to choose a representative structure for this family solved with a very good resolution: The Anisotropic Refinement Of Kunitz Type Domain C5, a structural protein solved through X-Ray diffraction at 0.95 Angstrom (PDB id :1KTH).
Once we found a structure that represent the kunitz domain, we searched for structure that are similar; so we performed a multiple structural alignment with PDBe Fold (v2.59.14 Apr 2014) [15, 16].
PDBe Fold is online and allows the pairwise comparison and 3D alignment of protein structures using the secondary structure. The parameters used to perform the alignment are reported in Table 1.

QUERY	
SUBMISSION FORM	*Pairwise*
SOURCE	*PDB entry*
PDB CODE	*1kth*
CHAIN	*A*
Lowest acceptable match	*70%*

TARGET	
SOURCE	*Whole PDB archive*
Lowest acceptable match	*70%*

Table 1: PDBe Fold parameters for the pairwise structural alignment

PDBe Fold alignment has examined 146145 entries, (380728 chains)and found 557 matches (proteins with a similar shape to the one of our query 1KTH). Then we derived a list of all the common PDB ids that have been reported from both the PDB search results (173) and the PDBe fold alignment (577), and we ended up with a list of 166 protein . Here we have common structures that contain the Kunitz domain and are also very similar to the structure of a kunitz protein, furthermore, since they have the same shape they are more likely to perform the same function. From this list of 166 common PDB ids we retained just one chain for each structure and we ended up with 87 structures listed in Table S2 (supplementary materials).We derived the fasta sequence for each of the 87 structures using a Python script (Get fasta from a list of PDB ids, supplementary materials paragraph 3.1).
In order to remove the redundancy we used

Mutational Analysis of Human TFPI-2

Inhibitor	NH_2---P_6	P_5	P_4	P_3	P_2	P_1	P_1'	P_2'	P_3'	P_4'	P_5'---------P_{18}'- COOH
BPTI	--- Y	T	G	P	C	**K**	A	R	I	I	R---------F---
APPI	--- E	T	G	P	C	**R**	A	M	I	I	R---------F---
APPH	--- M	T	G	P	C	**R**	A	V	M	P	R---------F---
TFPI Domain 1	--- D	D	G	P	C	**K**	A	I	M	K	R---------F---
TFPI Domain 2	--- D	P	G	I	C	**R**	G	Y	I	T	R---------F---
TFPI Domain 3	--- D	L	G	L	C	**R**	A	N	E	N	R---------F---
TFPI-2 Domain 1	--- D	Y	G	P	C	**R**	A	L	L	L	R---------F---
TFPI-2 Domain 2	--- V	D	D	Q	C	**E**	G	S	T	E	K---------F---
TFPI-2 Domain 3	--- D	E	G	L	C	**S**	A	N	V	T	R---------F---
HAI-1 Domain 1	--- K	V	G	R	C	**R**	G	S	F	P	R---------F---
HAI-1 Domain 2	--- D	T	G	L	C	**K**	E	S	I	P	R---------F---
HAI-2 Domain 1	--- V	V	G	R	C	**R**	A	S	M	P	R---------F---
HAI-2 Domain 2	--- V	T	G	P	C	**R**	A	S	F	P	R---------F---
IαTI	--- S	A	G	P	C	**M**	G	M	T	S	R---------F---
PLI	--- Y	T	G	P	C	**K**	A	R	M	I	K---------F---
UPTI	--- Y	T	G	P	C	**R**	A	H	F	I	R---------F---
SPI 1	--- K	T	G	P	C	**K**	A	A	F	Q	R---------F---
AsKC-1	--- D	V	G	R	C	**R**	A	S	H	P	R---------F---

.png

Figure 3: Alligned kunitz domain sequence.

Blastclust (Oct 2011) [11] that collects together things under a specific threshold of:

- L = coverage
- S = sequence identity

We decided to remove the redundancy with a threshold of 95% of coverage and the 99% of sequence identity, so we adopted the syntax reported in supplementary materials (subsection 8.1). Blustclust produces a file as output of the clustering process; this file contains a set of lines and in each of these lines there are all the elements that belongs to the same cluster. Finally we sorted the structures belonging to each cluster for their resolution, in order to facilitate the choice of the best representative protein structures of each cluster. For the sorting we used a Python script reported in supplementary materials (3.2 Paragraph: Sort PDB ids for their resolution). In order to remove the redundancy all the clusters have been evaluated and it was selected just the best structure for each one, 3 different parameter moved the choice :

- The resolution of the structure
- The presence of mutations
- The length of the protein

So we derived a list of 18 structures with the Kunitz domain,with a similar shape to the one of 1KTH Kunitz protein, without mutation, with a good resolution and all approximately of the same length (reported in Table S3, Supplementary materials). We finally used these structures to perform a multiple structure alignment within PDBefold with the parameter reported in Table S4 (Supplementary materials). The resulting Fasta multiple structural alignment was downloaded and trimmed in order to eliminate the beginning and ending gaps (the result of the alignment was reported in 5 paragraph of Supplementary materials).
In order to test the predictors we created two different testing sets:

- A negative set composed by all the protein that do not have the Kunitz domain. We obtained the negative set of 549798 protein through the download of all the Uniprot ids of proteins that do not contain the Pfam pf00014 and with a minimum length of 40 amino acids.
- A positive set composed of 344 proteins that contains the domain but were not included in the training set we used for build the model. The Training and testing sets must be different also in terms of sequence similarity, otherwise they will lead to overfitting. We first mapped all the PDB ids used to perform the structural alignment within Uniprot in order to archieve all the Uniprot ids of the proteins used to build the model. Then we compared the list of

Uniprot ids containing the Kunitz domain with the list of mapped ids and we finally ended up with a list of uniprot identifiers composing our positive testing set.

2.2 Model generation

Once the multiple structure alignment have been produced, the aim is to find out the model that is more likely to generate those observed sequences. In order to generate a model given the sequences we used HMMER(v3.2.1) [12, 14], a software package for database searching using profile HMMs. It includes several programs such as:

- hmmbuild: build a profile HMM from a multiple sequence alignment.
- hmmsearch: search a sequence database with a profile HMM.

We used hmmbuild to produce a model (model.hmm) starting from an alignment given as input using the syntax reported in supplementary materials (subsection 8.2)

2.3 Prediction

We used a python script to obtain the fasta sequence of the protein ids contained both in the positive and negative sets. Once we obtained the positive testing set and the negative testing set in Fasta format we performed a hmmsearch in order to verify whether our model will be able to predict correctly the positive set as belonging to the Kunitz family and the negative set as do not belonging to the family. The hmmsearch have been performed with the syntax for the positive set reported in supplementary materials (subsection 8.3).
And it returned an output where for each protein sequence there were reported 3 different values: e-value, score and bias, for both the full sequence and the domain. We did the same for all the proteins belonging to the negative set, so we performed the hmm search with the syntax reported in supplementary materials (subsection 8.4) where the e values threshold were specified at 100000000 and 100000 respectively for the all sequence and the domains. The output of the hmmsearch was a file called negative.search with 262614 hits and also in this file for each protein sequence there were reported 3 different values: e-value, score and bias for both the full sequence and the domain. We parsed both the positive.search and the negative.search files in order to have just a file with all the sequences e both the 2 e-values (the first for the full sequence and the second for the domains) as reported in Table S4 and Table S5 of supplementary materials. Then we created a file (Allsets.txt) with all the hints from the positive set and the negative set with the respective E-value (of the sequence) normalized by the number of element of the sets. For the positive sets the E-value of each protein was normalized by 344, the negative set was normalized by 549798. Finally we associated class 1 to all the protein belonging to the positive set and class 0 to all the protein belonging to the negative one. In the Allset.txt we also added all the proteins included in the negative sets that were not scored by the Hmm search (287184 proteins). Since they are shown to be very different from our model of kunitz protein, we expected the software not to report them and associated to them e-values of 1 and class 0.

2.4 Measures of the performance

In order to measure the performance the method we used was based on the maximization of the Matthew correlation coefficient. We used a python script (performance.py reported in supplementary materials) to compute the confusion matrix and the performance of our model. The confusion matrix is a 2x2 matrix for calculating the performance of prediction methods as reported in Table 2 In order to evaluate the performance we compute:

- The accuracy, a measure of how many predictions are correct on the overall

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

- True positive rate, a measure of how many of the real examples are correctly predicted

$$TPR = \frac{TP}{TP + FN}$$

- Positive predictive value, a measure of how many of the positive predictions are correct

$$PPV = \frac{TP}{TP + FP}$$

- Matthews correlation coefficient, it assumes 0 value for random predictions, 1 for perfect predictions and -1 value for completely wrong predictions.

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

	Condition positive	Condition negative
Test outcome positive	True positive	False positive
Test outcome negative	False negative	True negative

Table 2: Confusion matrix.

3 Results

3.1 Kunitz HMM model

Once we build the HMM profile for Kunitz proteins we produced a logo in order to graphically and intuitively represent the model we generated, since HMM Logos can help to compare families visually. In order to visualize the generated model we used Skylign, a tool for creating logos representing both sequence alignments and profile hidden Markov models [1, 20]. The logo of our model is shown in Figure 4, here for each emitting state of the HMM profile, it displays a series of letters. The height of those letters is determined by the deviation of the position's letter emission frequencies from the background frequencies domain. [18].
The produced logo highlights the conservation of the cysteines that form disulphide bridges that are known to be essential for the stability of the protein.

3.2 Optimization and performance

We obtained different confusion matrices for different threshold we tested (reported in Table 3 with their respective performance values), so we find out the best performance with a threshold of 0.000001.

Threshold	MCC	FPR	TPR
1	0.041	0.0	1.0
0.1	0.130	0.0	1.0
0.01	0.407	0.0	1.0
0.001	0.827	0.0	1.0
10^{-4}	0.982	0.0	1.0
10^{-5}	0.995	0.0	1.0
10^{-6}	0.997	0.0	1.0
10^{-7}	0.995	0.002	0.997
10^{-8}	0.994	0.005	0.994
10^{-9}	0.995	0.005	0.994

Table 3: reports all the performances obtained thought different thresholds. 10^{-6} was selected as the best one since it produces a MCC equal to 0.997.

Threshold = 0.000001

$$CM_{1\times10^{-6}} = \begin{bmatrix} 549791.0 & 2.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$$ACC = 0.9999963645419232$$

$$MCC = 0.997103824320116$$

$$FPR = 0.0\ ,\ TPR = 1.0$$

Setting the threshold at 10^{-6} the model shows good sensitivity (TPR =1.0) and accuracy (ACC=0.999). Moreover the computed Matthews Correlation Coefficient for our model is near 1 (MCC=0.997) and the number of false positive rate (FPR) was computed at zero, meaning that the overall performance the method is good.
Despite the optimization, two false positive were detected: however after closer inspection we concluded that they were wrongly included in the negative dataset, since on Uniprot they are reported as kunitz protein but they lack the Pfam Identifier (PF00014) which was used to build the dataset. These two proteins are:

- Kunitz-type serine protease inhibitor BiKTI (Uniprot ID:G3LH89), a serine protease inhibitor that inhibits plasmin [10]
- Ornithodorin (Uniprot ID: P56409), a potent and highly selective thrombin inhibitor. [19]

In order to plot the results of the confusion matrix we used a python script (roc_curve in supplementary materials) to have a better understanding of the performance of our model.
The ROC curve reported (Receiver Operating Characteristic) displays the ability of the binary classifier. On each axes is represented the sensibility (TPR= true positive rate) and the specificity (FPR= false positive rate).

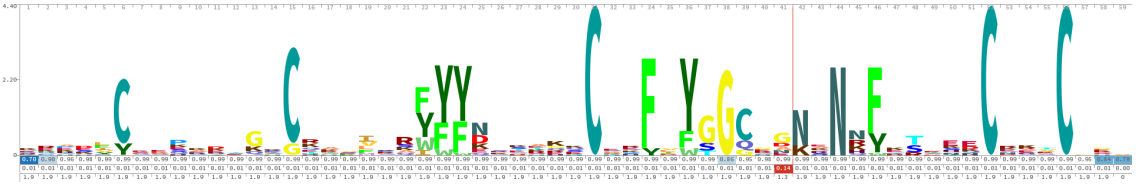

Figure 4: HMM logo of Kunitz protein.

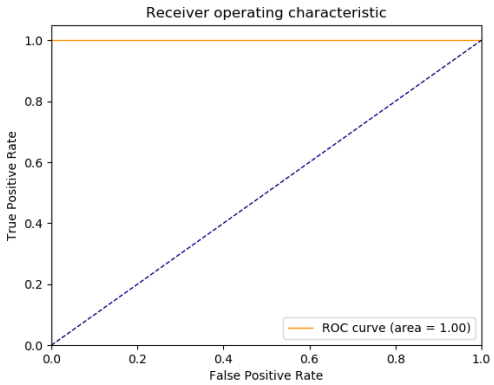

Figure 5: The Roc curve, on the x axis the false positive rate and on the y axis the true positive rate. The area under the curve is equal to 1 meaning that the model we generated is reliable.

The resulting curve shows immediately what is the dependency between these two measures in our data (Figure 5). The area under the curve was computed at 1 meaning that the model we generated is reliable.

4 Discussion

The use of profile hidden Markov models (HMM) has been widely adopted with the final aim of representing protein families. The method we performed consists on the use of few related sequences to build a profile HMM, which we then optimized using sequences annotated in Swissprot . We built a model of the kunitz domain that was able to perfectly detect the sequences belonging to the kunitz family and discard all the Non-Kunitz ones. Finally we evaluated the performance of our results in terms of specificity and sensibility of the model we produced. This analysis allowed us to affirm that the HMM profile of the kunitz domain is able to correctly perform an automatic classification of the Kunitz domain and therefore can be used for the annotation of uncharacterized domains.

References

[1] Interactive logos for alignments and profile HMMs. http://www.skylign.org/, 2019. [Online; accessed 13-April].

[2] worldwide PDB. www.wwpdb.org, 2019. [Online; accessed 13-April].

[3] W Antuch, P Güntert, M Billeter, T Hawthorne, H Grossenbacher, and K Wüthrich. Nmr solution structure of the recombinant tick anticoagulant protein (rtap), a factor xa inhibitor from the tick ornithodoros moubata. *FEBS letters*, 352(2):251–257, 1994.

[4] Bernadette Arnoux, Arnaud Ducruix, and Thierry Prangé. Anisotropic behaviour of the c-terminal kunitz-type domain of the $\alpha 3$ chain of human type vi collagen at atomic resolution (0.9 å). *Acta Crystallographica Section D: Biological Crystallography*, 58(7):1252–1254, 2002.

[5] AJ Barrett. An introduction to the proteinases. in "proteinase inhibitors"(aj barrett and g. salvesen, eds.), 1986.

[6] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[7] Stephen K Burley, Helen M Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M Duarte, Shuchismita Dutta, et al. Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research*, 47(D1):D464–D474, 2018.

[8] Gabriela Cesarman-Maus and Katherine A Hajjar. Molecular mechanisms of fibrinolysis. *British journal of haematology*, 129(3):307–321, 2005.

[9] Hitendra S Chand, Amy E Schmidt, S Paul Bajaj, and Walter Kisiel. Structure-function analysis of the reactive site in the first kunitz-type domain of human tissue factor pathway inhibitor-2. *Journal of Biological Chemistry*, 279(17):17500–17507, 2004.

[10] Young Moo Choo, Kwang Sik Lee, Hyung Joo Yoon, Yuling Qiu, Hu Wan, Mi Ri Sohn, Hung Dae Sohn, and Byung Rae Jin. Antifibrinolytic role of a bee venom serine protease inhibitor that acts as a plasmin inhibitor. *PLoS One*, 7(2):e32269, 2012.

[11] I Dondoshansky and Y Wolf. Blastclust (ncbi software development toolkit). *NCBI, Bethesda, Md*, 14, 2002.

[12] Sean Eddy. Hmmer user's guide. *Department of Genetics, Washington University School of Medicine*, 2(1):13, 1992.

[13] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.

[14] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[15] E Krissinel and K Henrick. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12):2256–2268, 2004.

[16] Evgeny Krissinel and Kim Henrick. Multiple alignment of protein structures in three dimensions. In *International Symposium on Computational Life Science*, pages 67–78. Springer, 2005.

[17] S Parkin, B Rupp, and H Hope. Structure of bovine pancreatic trypsin inhibitor at 125 k definition of carboxyl-terminal residues gly57 and ala58. *Acta Crystallographica Section D: Biological Crystallography*, 52(1):18–29, 1996.

[18] Benjamin Schuster-Böckler, Jörg Schultz, and Sven Rahmann. Hmm logos for visualization of protein families. *BMC bioinformatics*, 5(1):7, 2004.

[19] Andreas van de Locht, Milton T Stubbs, Wolfram Bode, Thomas Friedrich, Claus Bollschweiler, Wolfgang Höffken, and Robert Huber. The ornithodorin-thrombin crystal structure, a key to the tap enigma? *The EMBO journal*, 15(22):6011–6017, 1996.

[20] Travis J Wheeler, Jody Clements, and Robert D Finn. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden markov models. *BMC bioinformatics*, 15(1):7, 2014.