Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 2 April 2015
Applications Note

OXFORD

Report Laboratory of Bioinformatics (module B)

Development of Profile Hidden Markov Model for Kunitz domain

Andrea Sambugaro [1]

[1] Department of Biotecnology, University of Bologna, Bologna, Italy

Abstract

Motivation: Kunitz domain proteins inhibit the function of protein degrading enzymes, specifically they are protease inhibitors that play a fundamental role in the reduction of bleeding during surgery and other biological processes. Bovive pancreatic trypsin inhibitor (BPTI) is the classic member and among the most widely studied for the development of new drugs. Kunitz-domain has a monomeric structure that folds into a stable tertiary structure containing 3 disulphide bridges, 2 antiparallel β-sheets and a C-terminal α-helix.
In order to automatically classify proteins containing the Kunitz domain, starting from a multiple structural alignment of a selected set of proteins, we developed a profile Hidden Markov Model (profile-HMM): the model has been built and statistically validated using distinct sets of data.
Results: Our profile-HMM turns out to be a binary classifier with excellent accuracy in identifying proteins with Kunitz-domain. The parameters that measure performance indicate an accuracy very close to 100% and Matthews Correlation Coefficient very close to the optimal value 1. These results also allow to detect problems that do not concern the model itself but the SwissProt annotation.

1 Introduction

Proteins that inhibit the function of protease are characterized by the presence of the Kunitz domain. A relative small active domain with about 50 to 60 amino acids and a molecular weight of 6 KDa. Kunitz-type domains are common functional and structural elements of protein degrading enzymes which are usually associated with inhibition of trypsine-like serine proteases [2]. This domain appears in multi-domains proteins or in a single domain proteins such as aprotinin (bovine pancreatic trypsine inhibitor, BPTI), the most representative member of the kunitz-type serine protease inhibitors and one of most thoroughly investigated of all proteins [14]. Aprotinin was used as a medication administered by injection to reduce bleeding during heart and liver surgery.
Characteristic feature of the structure are the presence of six cysteine residues, a short α helix and two antiparallel β strands; an exposed binding loop project away from the supporting scaffold [10] (Figure 1). All six cysteines (position 5,14,30,38,51 and 55) are oxidize and form disulphide bonds, which contribute to the overall stability of the protein structure. The long, basic lysine 15 side chain on the exposed loop binds very tightly in the specificity pocket at the active site of trypsin and inhibits its enzymatic action. Kunitz domains are stable and standalone peptides able to recognize specific protein structures and also work as a competitive protease inhibitors in their free form [10].
In this project starting from a selection of a representative set of proteins and the available structural information, we developed an Hidden Markov Model (HMM) to automatically annotate the Kunitz domain proteins in SwissProt [6].
The HMM parameters are estimated from the output of the structural alignment of the selected sequences. Given a profile-HMM that represents a biological sequence family, we can use it to search a sequence database to find additional homologues that belong to the same family [13].

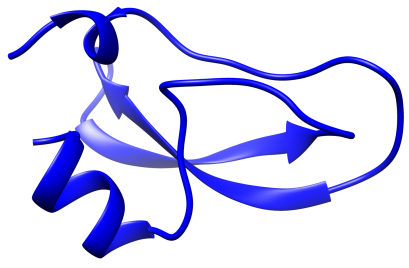

Fig. 1. Kunitz domain structure (bovine pancreatic trypsin inhibitor, PDB ID: 3TGI)[4].

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

2 Methods

2.1 Databases

The employed databases include Protein Data Bank (PDB, Update: May 8 2020)[7] to download the available 3D structure of Kunitz-type proteins (training set) in order to build the Hidden Markov Model (HMM). While from UniProtKB (Release: $2020-02$: $Apr22-2020$)[6], in particular SwissProt, are retrieved all the reviewed and manually annotated Kunitz-type and no-Kunitz-type proteins for building our validation (optimisation) and testing set for evaluating our model.

2.2 Data collection and processing

The model optimization and assessment are based on two distinct datasets: validation and testing set. Each set consists respectively of about the first half of a random set of both positives and negatives, and the second half. Specifically, the positive set consists of 353 proteins containing Kunitz domain. It was obtained removing the redundancy with respect to the proteins used for building the HMM (training set), starting from a data set of 359 entries retrieved from SwissProt, whereas the remaining ones, 561894 proteins, that do not contain the kunitz domain make up the negative set.

2.3 Computational methods

2.3.1 Structure selection and alignment

The selection of a representative set of kunitz-type structures from PDB taking into consideration the following parameters:

- Pfam accession number: PF00014 [8]
- number of polymer residues that range from 50 to 70
- resolution of the structure lower or equal to 3.5 Å

return 26 structures including the PDB ID: 2Y3K that contains more than one chain that was excluded. At the end we have 25 protein structures that have length around 60 residues and a molecular weight between 6.5-7 kDa.

To avoid redundancy in the proteins that will be our training set, through which the HMM can be build, we adopt a clustering procedure using BLASTClust 2.2.26 [5][3]. A BLAST package that allows us to cluster proteins according to personalized parameters. In our case setting the parameters for the clustering to 99% of sequence identity (option -S 99) over an area covering 90% of the length (-L 0.9), we obtained 14 clusters. For multiple clusters, only one structure was chosen as representative of the entire cluster according to the better resolution of the structure.

The structural alignment of the selected structures is performed on PDBeFold [9] multiple 3D alignment service, submitting the list of the 14 non redundant PDB ID. The result of this step shows that there was a discrepancy between the structure with the PDB code 2FMA, the one that represent a cluster, with respect to all the others. In particular, the Root Mean Square Deviation (RMSD) between this structure and all the others were higher than 2 Angstrom: for that reason, also the 2FMA structure was discarded. We finally obtained an alignment of 13 structure, PDB code and chain reported in Table 1, that are the ones selected for build up the HMM of the Kunitz domain.

PDB code: chain				
5PTI:A	1G6X:A	1KTH:A	6Q61:A	3OFW:A
5YV7:A	1DTX:A	1BPT:A	1BTI:A	1FAN:A
1NAG:A	7PTI:A	8PTI:A		

Table 1. Table shows the 13 selected PDB structures and the corresponding chain (PDB id: chain).

The output of the structural alignment is a FASTA file that will be the input for the generation of the kunitz domain model.

2.3.2 HMM generation

The HMM is generated with HMMER [1], hmmbuild method that reads a multiple sequence alignment, creates a new profile HMM, and save the result in a new hmmfile. In order to investigate the major characteristics of the model we used the web tool Skylign [12] that creates the logos of the profile for a better visualization. As we can see in Figure 2 the length between the two extreme cysteines is 50 residues; there are the six conserved cysteines involved in disulphide bridges fundamental for the correct folding of the domain.

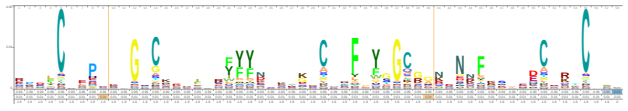

Fig. 2. HMM profile logo of the model: the logo shows the conserved position, in particular Cysteines

2.3.3 Testing and performance of the method

Once built the HMM model, the next steps are the optimisation, the assessment and the measure of the performance of the method. For this purpose is necessary to retrieve an opportune dataset that should be divided into two distinct subsets: the positive set containing the Kuitz-type protein domain and the other one, the negative set, composed of proteins which do not contain the Kunitz domain retrieved from UniprotKB, specifically from SwissProt, the manually annotated and reviewed section of UniprotKB database.

In order to obtain a fair test for our HMM of kunitz domain, we removed from the positive set the sequences that are too similar - share high sequence identity - to the ones that have been previously collected for generating the model. Therefore the NCBI distribution of blastpgp [3], a specialized protein blast comparison program, allowed us to exclude from the positives sequences with sequence identity higher or equal to 95%.

The two fold cross validation of the model was adopted, so both sets are divided into half to build the validation (random selection from positive and negative set) and testing set (the remaining ones) respectively.

The following step is the assessment of the dataset against the model. It is performed with the hmmsearch that given in input a profile Hidden Markov Model as a query, search against the database for significantly similar sequence matches. Against our positive and negative set the hmmsearch algorithm return a list of sequence hits and domain scores ranked according to the statistical significance (sorted by E-value). Some options need to be specified: –max that prevent cutting of distantly related proteins and increase the sensitivity and -Z 1 option that is important for normalising the E-value output. We point out that using the default E-value threshold (10) some sequences from the negative are filtered out. In order to measure the performance of the method, we had to reinclude them. In addition, the proteins belonging to the negative set are labelled with "0" while the positives with "1".

The basic performance measures are derived from the confusion matrix calculated using a program that takes in input the entries classified with the correct labels, "0" and "1" (observed labels), the correspondent predicted single best domain E-value and a given threshold. These observed labels are used to compared with the predicted labels for the performance evaluation after classification. The program returns a two by two table that contains four outcomes produced by a binary (two-class) classifier.

Various measures such as the accuracy (ACC)(1), the Matthews Correlation Coefficient (MCC)(2) as well as the sensitivity (SN)(3) and

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

the false positive rate (FPR)(4) have been derived for evaluate our HMM.

$$\mathrm{A}CC = \frac{(TP + TN)}{(TP + FN + TN + FP)} \tag{1}$$

$$\mathrm{M}CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{2}$$

$$\mathrm{S}N = \frac{TP}{TP + FN} \tag{3}$$

$$\mathrm{F}PR = \frac{FP}{TN + FP} \tag{4}$$

3 Results

Overall, our profile-HMM for the Kunitz domain appears to be a good binary classifier in distinguishing the proteins that contain the Kunitz domain from the others. Performance measurements such as Matthews correlation coefficient and accuracy, calculated from different levels of E-value threshold, allow us to find the optimal threshold value around $\dot{1}0^{-9}$ for the classifier.

The results obtained by applying our model to the validation set show good ACC and MCC values for an E-value threshold equal to $\dot{1}0^{-9}$: 0.999 and 0.997 respectively (Table 2). In addition, as can be seen from the corresponding confusion matrix (Table 3), there is only one wrong prediction, one false negative (Uniprot id: D3GGZ8).

Validation set		
Threshold	Accuracy	Matthew Cor. Coeff.
TH: 0.1	ACC: 0.96314	MCC: 0.12729
TH: 1e-05	ACC: 0.99998	MCC: 0.98877
TH: 1e-09	ACC: 0.99999	MCC: 0.99716
TH: 1e-14	ACC: 0.99998	MCC: 0.98862
TH: 1e-19	ACC: 0.99992	MCC: 0.94177

Table 2. Table shows the accuracy and Matthews correlation coefficient values for different threshold performed on the Validation set. The cells highlighted in yellow correspond to the E-value threshold of $\dot{1}0^{-9}$(1e-09) associated to the best ACC and MCC values.

		Prediction outcome	
		P′	**N′**
Actual value	**P**	True positive: **176**	False negative: **1**
	N	False positive: **0**	True negative: **280455**

Table 3. Confusion matrix Validation set. "P" positive, "N" negative,"′" means predicted.

In the event that several thresholds matched the best values of ACC and MCC, the optimal threshold was considered to be the one corresponding to the average of the optimal thresholds. Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset while the Matthew correlation coefficient a measure of the quality of binary classifications.

To further evaluate the correctness of the predictor we tested the performance compared to another set of proteins, the testing set. Applying the optimal threshold found for the Validation set ($\dot{1}0^{-9}$) to the Testing set, good results (Table 4) were observed as well, despite in addition to having a false negative (Uniprot id: O62247) there is also a false positive (Uniprot id: G3LH39)(Table 5).

Testing set		
Threshold	Accuracy	Matthew Cor. Coeff.
TH: 0.1	ACC: 0.96082	MCC: 0.12282
TH: 1e-05	ACC: 0.99996	MCC: 0.97535
TH: 1e-09	ACC: 0.99999	MCC: 0.99431
TH: 1e-14	ACC: 0.99997	MCC: 0.98281
TH: 1e-19	ACC: 0.99989	MCC: 0.91076

Table 4. Table shows the accuracy and Matthews correlation coefficient values for different threshold performed on the Testing set. The cells highlighted in yellow correspond to the E-value threshold of $\dot{1}0^{-9}$(1e-09) associated to the best ACC and MCC values.

		Prediction outcome	
		P′	**N′**
Actual value	**P**	True positive: **175**	False negative: **1**
	N	False positive: **1**	True negative: **281438**

Table 5. Confusion matrix Testing set. "P" positive, "N" negative,"′" means predicted.

The optimal threshold has been calculated not only by looking at the average value, but also by testing our model at different threshold for both validation and testing set. First the optimal threshold value was selected by using performance based on the validation set and testing it on the testing set. Later the same procedure was repeated but inverting the two sets. Both the procedures return as optimal threshold the one corresponding to the value of $\dot{1}0^{-9}$.

To support our results we also report the Receiver Operating Characteristic (ROC) curve of the model (Figure 3). The ROC plot is a measure that is based on two evaluation parameters, True Positive Rate (TFR, sensitivity) and False Positive Rate (FPR). For the confusion matrix of each E-value threshold are calculated the TFR, FPR and finally their are plotted in the ROC curve. Higher is the area under the curve, better is the performance of the method: as can be seen in Figure 3 the performance of our model is very high.

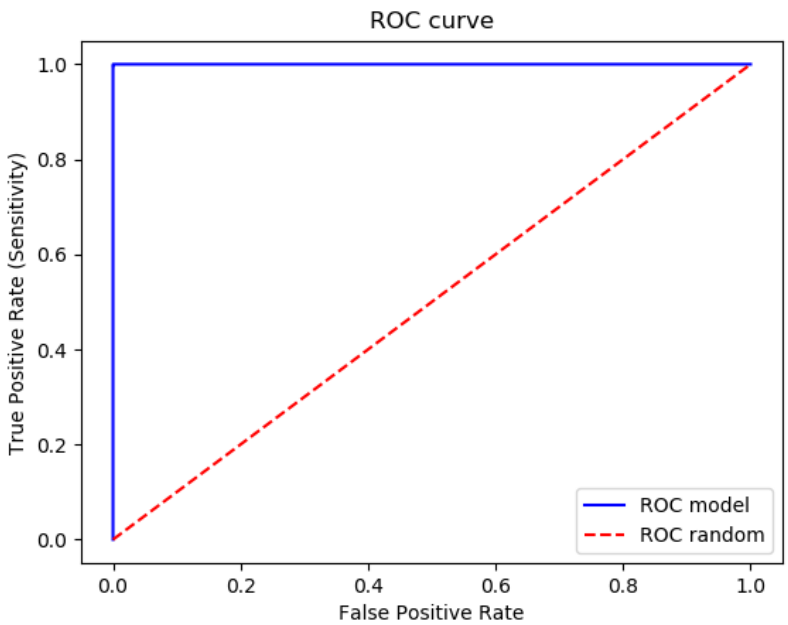

Fig. 3. ROC curve of the model

4 Discussion

In this paper we build a profile-HMM for the Kunitz-type domain. Screening our model on different datasets -validation and testing set- we found that the performance of the classifier to distinguish proteins containing the Kunitz domain against all the reviewed proteins in the SwissProt database is very good. Nevertheless, we can see that False Negatives and False Positives are predicted (Table 3, Table 5).

The False Negatives result to be the proteins with the Uniprot id D3GG78 and O62247 in the Validation and Testing set, respectively. Both of them are proteins encoded by the bli-5 gene in nematodes (*Haemonchus contortus*, *Caenorhabditis elegans*).

The alignment of the Kunitz domain of BLI-5 protein with bovine pancreatic trypsin inhibitor indicates a low sequence identity. In particular, in the BLI-5 Kunitz domain there is an absence of key residues of the BPTI motif. The sequence identity between this domain of each of the nematode species and BPTI is only approximately 20% [11]. This can be due to the long evolutionary distance between the Kunitz domain of nematode and the BPTI Kunitz domain but may also explain the absence of the serine protease inhibitory activity for the BLI-5 proteins.

Regarding the False Positives, only one protein is identify from the Testing set, G3LH39 (Uniprot id)(Table 5). After a thorough analysis in Uniprotkb, it appears that this protein despite containing the domain Kunitz is annotated without any Pfam domain referring to it. It is therefore an inconsistency in the database and not a prediction error of our model: the predicted single best domain E-value for G3LH39 is 4.2 e-23, a very high significant value, it is a Kunitz-type protein.

The profile-HMM highlighted also the conservation of the most important residues for the stability -six cysteine residues- and for the functionality -Lysin/Arginin in position 15- of the Kunitz domain among the family.

References

[1]hmmer.org, version HMMER 3.3.
[2]https://en.wikipedia.org/wiki/Aprotinin.
[3]https://launchpad.net/ubuntu/bionic/+package/ncbi-blast+-legacy.
[4]https://www.rcsb.org/structure/3TGI.
[5]S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
[6]A. Bateman. Corrigendum UniProt: the universal protein knowledgebase The UniProt Consortium. *Nucleic Acids Research*, 46(5):45, 2018.
[7]H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. Technical Report 1, 2000.
[8]S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, and R. D. Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47:427–432, 2018.
[9]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 I):2256–2268, 2004.
[10]S. Macedo-Ribeiro, C. Almeida, B. M. Calisto, T. Friedrich, R. Mentele, J. Stürzebecher, P. Fuentes-Prior, and P. J. B. Pereira. Isolation, cloning and structural characterization of boophilin, a multifunctional Kunitz-type proteinase inhibitor from the cattle tick. *PLoS ONE*, 3(2):1–17, 2008.
[11]G. Stepek, G. McCormack, and A. P. Page. The kunitz domain protein BLI-5 plays a functionally conserved role in cuticle formation in a diverse range of nematodes. *Molecular and Biochemical Parasitology*, 169(1):1–11, jan 2010.
[12]T. J. Wheeler, J. Clements, and R. D. Finn. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1):7, jan 2014.
[13]B.-J. Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6):402–415, 2009.
[14]M. Zweckstetter, M. Czisch, U. Mayer, M. L. Chu, W. Zinth, R. Timpl, and T. A. Holak. Structure and multiple conformations of the Kunitz-type domain from human type VI collagen α3(VI) chain in solution. *Structure*, 4(2):195–209, 1996.