Building a Profile Hidden Markov Model of the Kunitz-type domain

Alessandro Caula
Bioinformatics Master's Degree Course, University of Bologna, Bologna Italy

Abstract

Motivation*:* The aim of the project was to build a profile HMM based on structural alignment through which automatically classify the Kunitz-type protein domain.
The pancreatic Kunitz inhibitor, also known as aprotinin, bovine basic pancreatic trypsin inhibitor (BPTI) is a widely studied globular proteins. Beside a lot of major roles in which these proteins are involved, attracts a lot of interest the action of the BPTI when used in selected surgical interventions as it significantly reduces haemorrhagic complication and thus blood-transfusion requirements. Chosen a set of representative structures of the Kunitz-type domain the profile Hidden Markov model (HMM) of such domain was built and then statistically validated employing the 2-fold cross-validation method upon different set of data.
Result*:* The analysis of the performances of the model at different e-value threshold denoted that the model worked with an accuracy of approximately 100% and a Matthews correlation coefficient equal or very close to 1 in most the cases resulting in being a reliable predictor for the Kunitz-type domain.

1. Introduction

Kunitz domains are the active sites of proteins which inhibit the function of protein degrading enzymes, or more precisely, the Kunitz-type domains act as protease inhibitors.
With an average length of about 50 to 60 amino acids and a molecular weight of 6 kDa they are considered as relatively small molecules.
Among different example of Kunitz-type proteases inhibitors, the bovine pancreatic trypsin inhibitors (BPTI) is an extensively and comprehensively studied model structure. It is a monomeric globular polypeptide derived from bovine lung tissue. It has a molecular weight of 6512 Da and its single chain is 58 residues long that folds in a stable and compact tertiary structure characterized by 3 disulphide bridges, a twisted β-hairpin and a C-terminal α-helix.
The 3 disulphide bonds linking the 6 cysteine members of the chain (Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51) are responsible for the high stability of the molecule, while the protease inhibition action is carried out by the long basic lysine 15 (Lys-15) residue on the exposed loop that binds very tightly in the specific pocket at the active site of trypsin, inhibiting its enzymatic action. The conservation of the Lys/Arg 15 and Cys residues hold true for many BPTI-like inhibitors[1] (**Figure 1**).
Starting from available structural information of the Kunitz domain it is possible to generate a Hidden Markov Model (HMM) in order to automatically annotate the presence of the Kunitz-type domain with a certain reliability.
Hidden Markov Models are applied to the problems of statistical modelling, database searching and multiple sequence alignment of protein families and protein domains.[2]

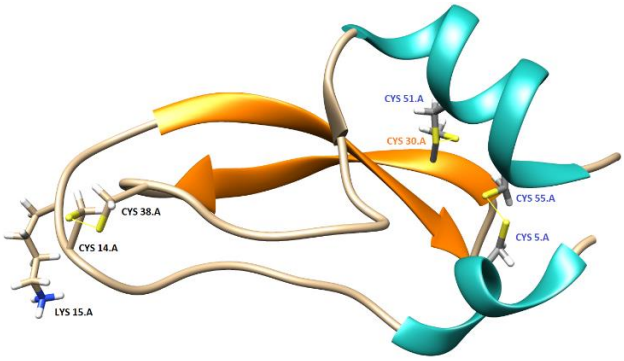

Figure 1. Structure of the BPTI Kunitz domain (PDB ID 5PTI). Highlighting the disulphide bridges, the twisted β-hairpin (orange) and the C and N-terminal α-helixes (blue).

Profile HMMs turn a multiple sequence alignment of a particular domain family into a position-specific scoring system suitable for searching database for homologous sequences and for representing sequence families.[3]
When employed in discrimination tests the HMM profile is able to distinguish members of these families from non-members with a high degree of accuracy.

2. Methods

2.1 Databases

The sequences of the structurally resolved Kunitz-type proteins used to construct the HMM model are retrievable in Protein Data Bank (PDB) database[4] (release 2020_05).
In UniProtKB[5] (release 2020_02) are found all the sequences, information on the functions and important sites of the Kunitz-type and non-Kunitz type proteins employed for the optimization and validation process.
In Pfam[6] (release 2018_09) there are detailed descriptions of the Kunitz BPTI family and other information on this domain.

2.2 Computational methods

BLAST+[7] version 2.2.26 was used for the clusterization process of the protein sequences and for running database searches in our own server without size, volume and database restrictions.
PDBeFold[8] (release 2014_04) it's an interactive on-line service utilized for the multiple structural alignment of the representative proteins then used to build the profile HMM.
HMMER[9] version 3.2.1 (release 2018_06) was employed for the construction and the research of the profile HMM against the protein sequence databases.
Skylign[10] web service has been used for the creation of the logo representation of the Hidden Markov Model profile.

2.3 Workflow

2.3.1 Representative Structures selection

For the selection of the reference set of proteins then used for the construction of the profile HMM, has been performed an advance search on PDB database. Particularly looking for those proteins annotated in the Pfam database with the Kunitz-type domain (Pfam accession number: PF00014) that have a total number of polymer residues between 50 and 70, and whose structure has been resolved with a resolution lower than 3.5 Å. From this query a set of 25 protein structures were obtained and then downloaded in the FASTA format.
In order to avoid the problem of redundancy in the final reference set, the next step was to implement a clusterization procedure using the *blastclust* program of the BLAST+ package. By specifying the length of the coverage at 0.95 and the score of the similarity at 0.99, the program begins with pairwise matches and places in the same cluster all the sequences that share a degree of sequence similarity of at least 99% within a coverage length no lower than the 95%.
To avoid reference set abundance, from the resulting 14 clusters only the best resolved protein structure within the clusters were chosen as the representative one of the clusters. Moreover, for certainty was manually verified the actual presence of the Kunitz-like domain for these set of proteins.

2.3.2 Protein alignment

Before building the profile HMM the multiple structural alignment of the representative set of proteins has been performed with PDBeFold multiple structural alignment service (**Figure S1**).
The first and second results of the alignment highlighted how the PDB ID proteins 2FMA and 8PTI had an overall higher Root Mean Square Deviation (*2.3556* and *1.2516* respectively) and a lower Q-score of *0.2097* and *0.3354* in respect to the other aligned proteins. These two proteins were therefore filtered out from the reference set

of proteins. Once satisfied with the results of the third attempt the FASTA format of the alignment were downloaded. This will be the input file for the profile HMM building procedure.

2.3.3 Modeling the Kunitz-domain with an HMM

The profile HMM was generated by applying the *hmmbuild* function of HMMER software. This method by reading as input the multiple sequence alignment gained in the previous step builds a new HMM model and saves it in an HMM format file.
After the generation of the model the Skylign tool has been used to generate the logo of the model, this graphical representation highlighted the conservation of the six Cysteines as well as for the Tyrosine in our profile HMM (**Figure S2**)

2.3.4 Building the benchmark sets

The adopted sets of proteins used for the evaluation of the model binary classification ability were collected from UniProtKB.
For the positive set all the proteins annotated with the Kunitz-like domain were downloaded in a resulting set of 359 protein sequences. In order to avoid redundancy problems, using the *blastpgp* program of the BLAST+ package all the sequences that shared a degree of sequence identity greater than 90% with the representative proteins used to build the profile HMM were filtered out, ending up with a final positive dataset of 350 protein sequences.
Whereas, for the negative set a total of 561894 proteins which do not present the Kunitz domain were downloaded. No further modifications were executed on this set.

2.3.5 Model training and validation

For the purpose of the 2-fold cross validation technique used to evaluate the predictive capacity of profile HMM, the positive and negative datasets were randomly shuffled and then partitioned into training and test sets of equal size. The first consisting by the merged first halves of the positive and negative sets while the test by the second halves (**Table 1**).

	Training set	Testing set
Positives	175	175
Negatives	280'947	280'947

Table 1. Number of protein sequences in the training and test sets after the partition of the positives and negatives benchmark sets.

Once the datasets were split into two equal parts the *hmmsearch* program of the HMMER package, has been used to search the profile HMM against the positive and negative databases and score the alignments.
By specifying the *"--max"* and the *"-Z"* options of the hmmsearch procedure, respectively the heuristics for cutting of the distantly related proteins was turned off and the normalization of the e-value output was set to 1.
The outputs of this method consist in a list of the best scoring sequences and domains ranked by their statistical significance (e-value).
For the classification purposes, all the distantly related proteins from the negative subsets for which the hmmsearch didn't return any match, were reinserted and assigned with an e-value equal to the default one of the operation (*10.0*).

2.3.6 Performance evaluation

When all the hmmsearch results for the subsets were collected in tabular files containing protein ID, e-value and protein class (0 for the positive and 1 for the negatives) an in-house python script have been used to optimize the classification e-value threshold on the training subsets and then validate the performance on the test sets. The parameter that the script computed were the Accuracy, the Matthews correlation coefficient, the True and False Positive Rate and the construction of the confusion matrix of the HMM prediction.

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

Matthews correlation coefficient (MCC):

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

3. Results and Discussions

The analysis of the Accuracy and the Matthews correlation coefficient confirmed the profile HMM as a good predictor for the Kunitz-type proteins classification.

The first application of the performance of the model has been done on the training sets at different e-value threshold, finding as the optimal thresholds the range from *1,00E-08* to *1,00E-11*, showing a perfect prediction of the presence of the Kunitz domain with an accuracy and a MCC of 1.00. Indeed, in this range of e-values no False Negatives or Positives were found in the confusion matrices.

The fairness of the model predictions has been then validated on the testing sets using as threshold the e-values range obtained from the training procedure. As shown in the **Table 2** also in this case the results were satisfying in particular at an e-value threshold of *1,00E-08* and *1,00E-09* but with some proteins recognized as False Negative and False Positives (**Table 3**).

The choice of the best threshold for which the model preforms in an optimal way the binary classification fell on the lowest e-value (*1E-09*).

Threshold	ACC	MCC	TPR	FPR
1,00E-08	0,999986	0,9914027	0.98857	3,60E-06
1,00E-09	0,9999893	0,9914027	0.98857	3,60E-06
1,00E-10	0,9999858	0,9885149	0.98286	3,60E-06
1,00E-11	0,9999858	0,9885149	0.98286	3,60E-06

Table 2. Results of the testing sets performances within the optimal range (1,00E-08 to 1,00E-11) obtained from the performance of the training set.

		Actual class	
		Kunitz-type	Non Kunitz-type
Predicted class	Kunitz-type	**173** (TP)	1 (FP)
	Non Kunitz-type	2 (FN)	**280'946** (TN)

Table 3. Confusion matrix of the HMM prediction of the Kunitz and non-Kunitz type domain at the threshold of 1,00E-09. The matrix highlights the presence of one False Positive (FP) and two False Negatives (FN).

A further evaluation of the diagnostic ability of the model is obtained by computing the ROC curve. This is a graphical plot created by plotting the True Positive Rate against the False Positive Rate at various e-values threshold settings.

True Positive Rate (TPR) or sensitivity:

$$TPR = \frac{TruePositives}{TruePositives + FalseNegatives}$$

False Positive Rate (FPR):

$$FPR = \frac{FalsePositives}{TrueNegatives + FalsePositives}$$

The curve tells how much the model is capable of distinguishing between the two classes (positive and negatives).

The higher is the area under the curve (AUC), the better is the model in terms of precision and accuracy. In particular, the ROC curve plotted with our model performances has an AUC equal to 1, confirming once again the accuracy of the model prediction (**Figure S3)**.

The model has been then cross validated by evaluating the performance first on the testing set and then validating the best e-value threshold on the training set, the results corroborate the optimum e-value at 1,00E-09.

A more in deep reasoning has been done on those proteins from the validation set that were identified as false positives and false negatives at the e-value threshold of 1,00E-09. After retrieving the identification number of the false positive protein (*UniProt ID: G3LH89)* and a further analysis of such protein we can conclude that the protein presents the Kunitz-type domain but that it is not annotated with the Pfam domain referring to it in the UniProtKB database.

While for the false negative proteins (*UniProt ID: O62247 and D3GGZ8*), for which the hmmsearch return an e-value of 7.70E-06 and 9.60E-05 respectively, they do are annotated with the Kunitz-type domain in Pfam but for both the proteins, the protease inhibitor activity is uncertain since it seems that they lack in some features of the Kunitz-type proteins. Moreover, the D3GGZ8 doesn't have a high score of

annotation but its function has been inferred by homology. Both these sequences share a low sequence identity with our set of representative proteins used to build the HMM and this can also be due to their evolutionary distance.

References

1. Ascenzi, P., Bocedi, A., Bolognesi, M., Spallarossa, A., Coletta, M., Cristofaro, R. D., & Menegatti, E. (2003). The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. *Current Protein and Peptide Science*, *4*(3), 231-251.

2. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology*, *235*(5), 1501-1531.

3. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics(Oxford, England)*, *14*(9), 755-763.

4. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., ... & Fagan, P. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, *58*(6), 899-907. (https://www.rcsb.org/)

5. UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic acids research*, *43*(D1), D204-D212. (https://www.uniprot.org/)

6. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... & Sonnhammer, E. L. L. (2019). The Pfam protein families database in 2019. *Nucleic acids research*, *47*(D1), D427-D432. (https://pfam.xfam.org/)

7. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1), 421.

8. Krissinel, E., & Henrick, K. (2004). 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. (https://www.ebi.ac.uk/msd-srv/ssm/)

9. Eddy, S. (1992). HMMER user's guide. *Department of Genetics, Washington University School of Medicine*, *2*(1), 13.

10. Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC bioinformatics*, *15*(1), 7.

Supplementary Material

Alessandro Caula

Some files as the Hidden Markov Model, the outputs of the hhmsearch and the alignment as well as for the python script that have been used are found on this repository on GitHub: https://github.com/AlessandroCaula/LaboratoryOfBioinformatics_profile-HMM

Supplementary Images and plots

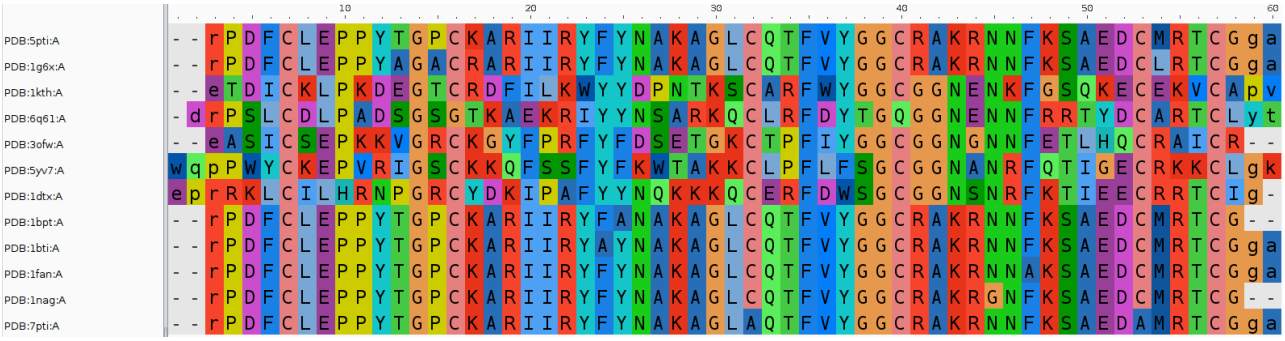

Figure S1. AliView image of the multiple structural alignment of the reference set of proteins performed by PDBeFold.

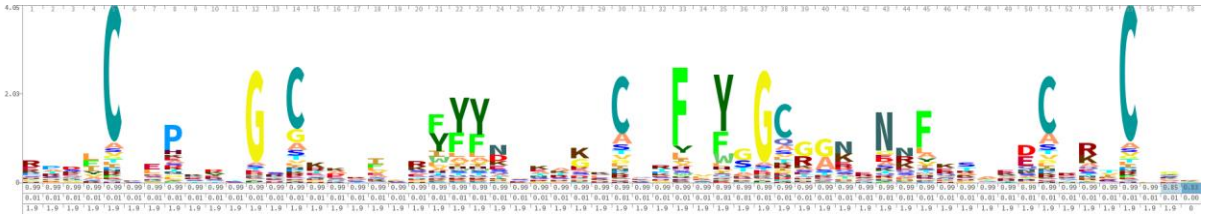

Figure S2. Logo representation of the profile HMM using Skylign web service.

Supplementary Images and plots

Alessandro Caula

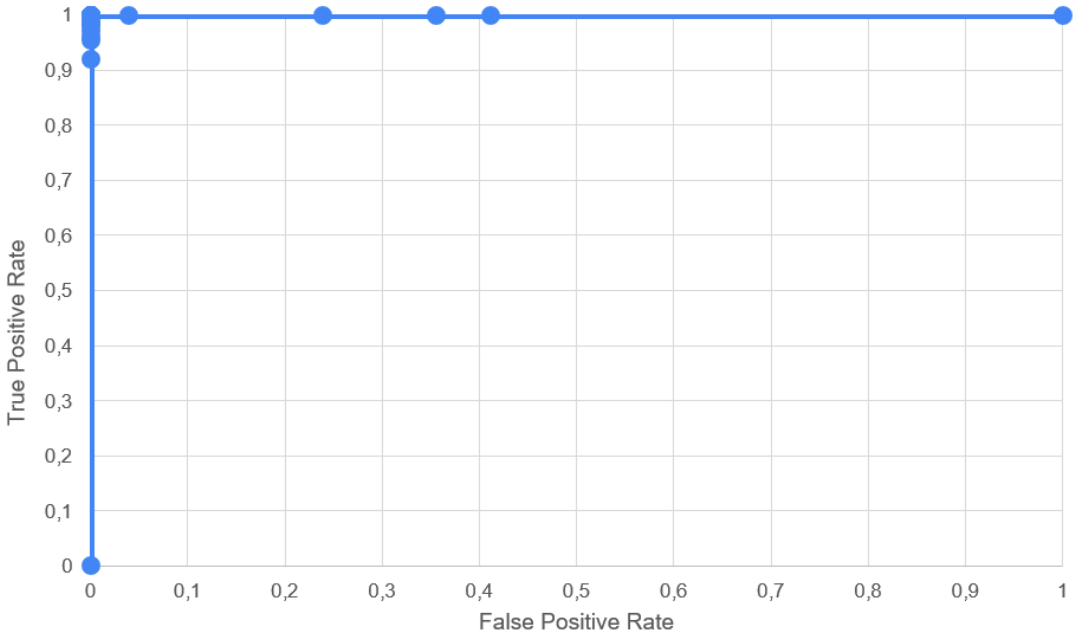

Figure S3. ROC curve of the classification model performance at different e-value thresholds. where the False Positive Rate and True Negative Rate are plotted together. The Area Under the Curve is equal to 1.