Bioinformatics

Building a profile Hidden Markov Model of the Kunitz BPTI domain for annotating proteins in UniProt/SwissProt

Lorenzo Campini[1,*]

[1]Bioinformatics Master, Department of Pharmacy and Biotechnology, University of Bologna

Abstract

Motivation: In this project, the aim is to develop a Hidden Markov Model (HMM) that enables accurate identification and annotation of the Kunitz BPTI domain (Pfam ID = PF00014) on UniProt SwissProt, starting from Multiple Structure Alignments of a training set of PDB entries having the structure of the Kunitz domain.

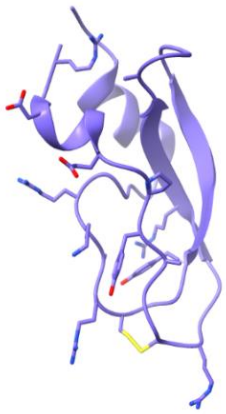
Results: Three HMMs were developed and were found to be with almost identical performances, all of them almost perfect classifier. Their performances were assessed through Matthew Correlation Coefficient and the Area Under the roc Curve values. Finally, thanks to this work some issues in the UniProt annotation method could be found.

Supplementary information: Supplementary data are available at https://github.com/Lorenzo-Campini/LB1_Kunitz_domain.

1 Introduction

Protease inhibitors play a critical role in regulating protease activity and are involved in various biological processes across different organisms. Among them, the Kunitz-type protease inhibitors have gathered significant attention due to their wide distribution and diverse functions [4-5].

Figure 1. The structure of the Kunitz BPTI domain



This is the structure of the Kunitz domain. At the bottom, the disulphide bridge and the Arginine of the active site can be seen.

The Kunitz BPTI domain, initially discovered in bovine pancreatic trypsin inhibitor (BPTI), represents a prominent example of this inhibitor family [6].

The Kunitz BPTI domain is a cysteine-rich peptide chain consisting of approximately 50 to 60 residues, its molecular weight is around 6 kDa. It is characterized by a conserved fold comprising two anti-parallel β-sheets and one or two helical regions (Figure 1). This domain's structure is maintained by three disulphide bridges at positions 2-52, 11-35, 27-48, as can be seen in the sequence logo from Pfam entry (Image 2); while its binding specificity is mainly determined by a solvent-exposed loop, encompassing residues 8 to 19 [1,2,3]. Within this loop, a highly exposed residue at position 15 (Image 1), typically arginine or lysine, plays a crucial role in inhibiting the activity of trypsin [6].

Examples of proteins containing the Kunitz BPTI domain include the Alzheimer's amyloid precursor protein (APP) and the tissue factor pathway inhibitor (TFPI), highlighting its significance beyond protease inhibition [1].

Moreover, Apoprotein can inhibit the the proteolytic activity of thrombin and plasmin, showing to reduce blood loss and blood requirement when administered prior to surgery [7].

To effectively identify and annotate the Kunitz BPTI domain, computational methods have been developed, such as Hidden Markov Models (HMMs), which are statistical models capable of capturing hidden information in observable sequencing data [8]. HMMs have been widely employed in bioinformatics, particularly for sequence analysis, motif detection, and protein classification [9]. By building an HMM for the Kunitz BPTI domain, we can leverage the model's ability to recognize conserved patterns and dependencies in amino acid sequences related to protein structures [11].

In this project, the pipeline follows some key steps. Starting from selecting representative protein structures containing the Kunitz domain, a multiple structural alignment (MSA) is then generated. Using this alignment, the construction of an HMM profile specific to the Kunitz BPTI domain is conducted. Finally, in order to optimize the E-value threshold, which is crucial to the classification, this parameter is fine-tuned using the k-fold cross validation technique on the UniProt/SwissProt database [11].

2 Methods

2.1 Selection of the structures

Three HMMs have been developed in this project, starting from MSAs from different training sets of sequences. In order to have a multiple structure alignment, the proteins were selected thanks to advanced search tool on PDB [12] as displayed in the table (Table 1), using a combination of Pfam [13], SCOP2 [14,15] and CATH identifiers [16]. The Entry ID and the Auth Asym ID, to identify the asymmetrical units, were downloaded as csv files (.csv files in supplementary data).

Table 1. Filters for the selection of the PDB structures

Model	Identifier(s)	Data collection Resolution	Polymer entity length	Polymer entity mutation count
First	Pfam = PF00014	< 3.0 Å	50<= <70	1
Second	Pfam = PF00014 SCOP2 = 4003337	< 3.0 Å	50<= <70	/
Third	Pfam = PF00014 CATH = 4.10.410.10	< 3.0 Å	50<= <70	/

This table indicates the different filters used to select the structures of the PDB entries used for training the HMM.

2.2 Generation of the HMMs

The MSAs were subsequently generated with the PDBeFold online tool [17], submitting the lists of PDB IDs.
Coherently to what are the positions of the Pfam model of the Kunitz BPTI domain, the different MSAs where trimmed one position upstream with respect to the first Cysteine and one position downstream with respect to the last Cysteine of the domain (.fasta files for each model in the supplementary data).
The HMMs were then built with HMMER [18]. This procedure was carried out thanks to the command with default options:

```
$ hmmbuild model.hmm clean_MSA.fasta
```

HMMER's hmmbuild read the MSAs and from those it creates an HMM (.hhm files in supplementary data) automatically trimming the nonsignificant positions (i.e., the positions mostly occupied by gaps in the MSA).
The hmm logos were then created with the command:

```
$ hmmlogo model.hmm
```

2.3 Selection of the testing set

Subsequently, from the PDB IDs, the corresponding UniProt IDs were downloaded thanks to the UniProt mapping tool.

The dataset used for testing has been retrieved from UniProt, using as negative testing examples those proteins that are not under the Pfam ID for Kunitz BPTI (PF00014), while using as positives those proteins that are annotated with that ID.
The proteins that were used to train the model have been removed from the positive testing set, alongside those proteins that had sequence identity higher or equal to 95% with the formers. This operation was carried out thanks to the blast+ [19] suite and these two commands:

```
$ makeblastdb -in training_kunitz.fasta -dbtype
"prot" -title training_db -out Tset.db

$ blastpgp -i partial_testing_kunitz.fasta -d
Tset.db -m 8 -o positives.bl8
```

With the first command a database blast-like was created and then, with the second command, the positive Kunitz examples were blasted against the database ("blasting" folder in the supplementary data).
Once the testing set IDs (positives and negatives) were decided, the whole testing set was downloaded from UniProt in fasta format.

2.4 Evaluation of the models

To evaluate the models, the command hmmsearch was used with the –noali and –max options in order to make the output file lighter and to turn off all the heuristics filters (.search files in supplementary data).

```
$ hmmsearch -Z 1 --noali --max
-o result.search model.hmm whole_test-
ing_set.fasta
```

The models were evaluated with a cross validation procedure, so the testing set was randomly divided into two batches. On one split the threshold optimization was performed, while on the other split the model with that threshold has been evaluated; and then vice-versa. The model's qualities were assessed with the Matthew Correlation Coefficient (Equation 1) and the F1 score (Equation 2).

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

Equation 1. The formula of the Matthew Correlation Coefficient is displayed in this equation.

$$F1 = 2 \cdot \frac{p \cdot r}{p + r}$$

Equation 2. The formula of the F1 score is displayed in this equation, with "p" being the Precision (TP/(TP+FP)) and "r" being the Recall (TP/(TP+FN))

3 Results and discussion

3.1 Model construction

From the search in PDB, a different number of structures were retrieved for each model (.pdbids in the supplementary data). Not all of them were suitable for the alignment, in particular 5NX1:D was too short to be a complete Kunitz domain (40 residues). Some PDB chain entries were not mapped to UniProt/SwissProt, that is because some PDB entries have the Swiss/Prot identifier in common, so that identifier was considered as representative.

22 sequences from the positive Kunitz testing set were found to have a sequence identity higher than 95% with respect to the sequences used in the training of the models, so they were removed from the testing set to avoid redundancy (redundancy.ids in /blasting in supplementary material).

3.2 Model evaluation

Given the fact that the testing set is a skewed class because the negative example are much more (three orders of magnitude) than the positive examples, the accuracy measure would have been biased for evaluating the models. For this reason, the MCC was used as a score to assess the quality of the model.

The confusion matrixes were computed using different thresholds for the classification, over the two batches per model (Tables 2-4) (\model_testing\model_batches in supplementary material). The average of the thresholds that gave the best MCC were then used to assess the final performance of the model (Table 3).

The assumption that the models evaluate just the Kunitz domain can be taken, that is because all the three HMM have 53 positions (just as much as the Pfam's HMM for the Kunitz domain) and a trimming of the MSA has been performed one position upstream the first Cysteine and one position downstream the last Cysteine (just like Pfam's HMM). Moreover as aforementioned, HMMER automatically does not consider the non-significant positions of the alignment. Additionally, only the monodomain chains for the Kunitz domain have been used for the training, thus reinforcing this assumption. Lastly, this is the reason why the e-values taken into consideration were the ones about the "best 1 domain", so to consider the best one local hit.

Table 2. Cross validation results for each batch of each model.

Models	Thresholds	MCC
First_model, Batch_1	$1{\cdot}10^{-8}$	1.0
	$1{\cdot}10^{-9}$	1.0
	$1{\cdot}10^{-10}$	1.0
	$1{\cdot}10^{-11}$	1.0
First_model, Batch_2	$1{\cdot}10^{-8}$	0.99465
	$1{\cdot}10^{-9}$	0.99732
	$1{\cdot}10^{-10}$	0.99463
Second_model, Batch_1	$1{\cdot}10^{-8}$	0.99686
	$1{\cdot}10^{-9}$	0.99686
	$1{\cdot}10^{-10}$	0.99372
Second_model, Batch_2	$1{\cdot}10^{-9}$	1.0
	$1{\cdot}10^{-10}$	1.0
	$1{\cdot}10^{-11}$	1.0
Third_model, Batch_1	$1{\cdot}10^{-8}$	1.0
	$1{\cdot}10^{-9}$	1.0
	$1{\cdot}10^{-10}$	0.99696
Third_model, Batch_2	$1{\cdot}10^{-8}$	0.99732
	$1{\cdot}10^{-9}$	0.99732
	$1{\cdot}10^{-10}$	0.99732
	$1{\cdot}10^{-11}$	0.99732

This is the table that recaps the different MCCs associated with the different thresholds when evaluating the model through a confusion matrix.

Table 3. Summary of the models' performances

Model	Best threshold	MCC	AUC
First	$1.88875{\cdot}10^{-9}$	0.99857	1.00
Second	$5.87000{\cdot}10^{-9}$	0.99857	1.00
Third	$4.13875{\cdot}10^{-9}$	0.99857	1.00

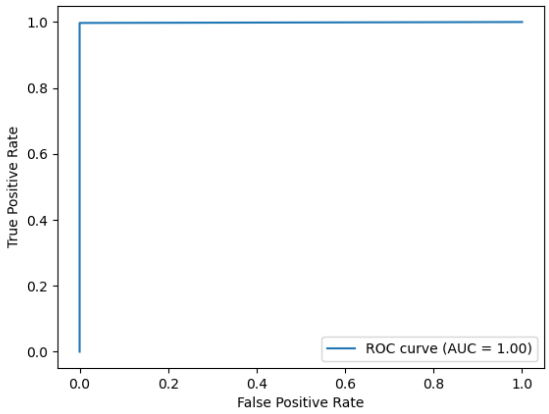
This table shows the best threshold, the MCC and the AUC for each model.

Table 4. Confusion matrix of the models

	Actual Positives	Actual Negatives
Predicted Positives	TP = 351	FP = 0
Predicted Negatives	FN = 1	TN = 569126

This is the confusion matrix, which is the exact same for all the models, computed with the best threshold for each model.

Figure 2. The models' ROC curve



The graph in the image represents the ROC curve of the models. The AUC is 1.00, thus indicating a perfect classifier.

An evaluation method used to visualize the performance of the models is the Receiver Operating Characteristic (ROC) plot. All the three models have the same ROC plot (Figure 2), and this curve indicates that these models are almost perfect in the predictions over the used dataset. The measure of the Area under the Curve (AUC) is 1.00 for all the three models, conversely to what would be a random classifier (AUC = 0.50) or a totally wrong classifier (AUC = 0.00)

As seen in the results, the three models are almost identical to each other's, but something can be said about their differences: the second one is indeed slightly more precise and has more recall than the other two models. It indeed tends to assign slightly higher e-values to the False Positives with comparison to the other two models and slightly lower e-values to the False Negatives with comparison to the other models (model_testing\model_considerations folder in the supplementary data). For this consideration, only the False Positives and False Negatives that are obtained with thresholds (range $1{\cdot}10^{-6}$-$1{\cdot}10^{-12}$) around the optimal ones are considered.

Moreover, the three different hmm logos (model_testing\model_considerations folder in the supplementary data) show that in the second model, the six positions (2-52, 11-35, 27-48) occupied by the Cysteines that are implied in the disulphide bridges, are more conserved than in the other two models, coherently to the hmm logo of Pfam. Nonetheless, the performances of the models are very similar to each other, so from now on only the second model will be taken into consideration for further analysis.

It is noteworthy to mention that the only false negative in the detection using the best threshold is the protein D3GGZ8. This protein is given an e-value of $7.7{\cdot}10^{-7}$ and is annotated in the SwissProt database with a score of 2/5 and it appears to lack serine protease inhibitor activity in vitro and all the catalytic features of serine proteases. Thus, this protein is classified as negative probably because it does not have all the characteristics in the sequence of the other proteins that have Kunitz domains, so further analysis should be performed on this protein to assess if it belongs to Kunitz BPTI or not.

The other false negatives that are detected at thresholds around $1{\cdot}10^{-12}$-$1{\cdot}10^{-13}$ are poorly annotated with scores 2/5 and functions inferred by similarity (Q11101, P86963, P0CH75), so the same can be applied as for as D3GGZ8. The entry P0CAR0 could potentially be a Kunitz BPTI domain, but it is too short in sequence to be detected as positive from the classifier, so it will be declared as negative even if it is just a fragment of Kunitz. The other two proteins (Q9BQY6, D9IFL3) are assigned rather low e-values (10^{-13}), and they are characterized from 4 or 5 disulphide bridges, instead of three; moreover, their functions are only electronically inferred.

Looking at the few false positives detected at thresholds around $1{\cdot}10^{-6}$, all of them do not have the Pfam ID, but are annotated as Kunitz by other methods, such as InterPro [20], SUPFAM [21-22] or PROSITE rules [23]. Thus, at slightly lower-than-optimal thresholds, the method is able to recognize proteins that contain the Kunitz domain but are note annotated by Pfam, CATH not SCOP2.

4 Conclusions

This work targets the development of a model for the identification of proteins that contain the Kunitz BPTI domain, following Pfam annotation. In order to carry out this task, three HMMs were build starting from separate sets of multiple structure alignment obtained PDB entries selected with different filters for each model.

As expected, the chosen HMM was able to correctly annotate all the proteins in UniProt/SwissProt database except from one false negative, of which the annotation score is quite poor, and the function is electronically inferred. The Matthew Correlation Coefficient and the Area Under the ROC Curve were used as evaluation methods for the model's performance, and they were respectively 0.99857 and 1.00, indicating an optimal classifier.

It is interesting to consider that the false positives that were detected at a threshold slightly higher than the optimal one are actually annotated as Kunitz BPTI from other methods other than those used for the obtaining the training set of sequences. On the other hand, the false negatives which was detected at thresholds slightly lower than optimal are poorly annotated and need further analysis to assess their molecular function.

Conflict of Interest: none declared.

5 References

1. Structure Mishra, M. Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. J Mol Evol 88, 537–548 (2020). https://doi.org/10.1007/s00239-020-09959-9

2. Structure Kassell, B., & Laskowski, M., Sr (1965). The basic trypsin inhibitor of bovine pancreas. V. The disulfide linkages. Biochemical and biophysical research communications, 20(4), 463–468. https://doi.org/10.1016/0006-291x(65)90601-7

3. Structure KASSELL, B., RADICEVIC, M., ANSFIELD, M. J., & LASKOWSKI, M., Sr (1965). THE BASIC TRYPSIN INHIBITOR OF BOVINE PANCREAS. IV. THE LINEAR SEQUENCE OF THE 58 AMINO ACIDS. Biochemical and biophysical research communications, 18, 255–258. https://doi.org/10.1016/0006-291x(65)90749-7

4. Function Medeiros, A. F., Costa, I. S., Carvalho, F. M. C., Kiyota, S., Souza, B. B. P., Sifuentes, D. N., Serquiz, R. P., Maciel, B. L. L., Uchôa, A. F., Santos, E. A. D., & Morais, A. H. A. (2018). Biochemical characterisation of a Kunitz-type inhibitor from Tamarindus indica L. seeds and its efficacy in reducing plasma leptin in an experimental model of obesity. Journal of enzyme inhibition and medicinal chemistry, 33(1), 334–348. https://doi.org/10.1080/14756366.2017.1419220

5. Function Kobayashi, H., Yagyu, T., Inagaki, K., Kondo, T., Suzuki, M., Kanayama, N., & Terao, T. (2004). Therapeutic efficacy of once-daily oral administration of a Kunitz-type protease inhibitor, bikunin, in a mouse model and in human cancer. Cancer, 100(4), 869–877. https://doi.org/10.1002/cncr.20034

6. Function Chang, Y. C., Wang, J. D., Hahn, R. A., Gordon, M. K., Joseph, L. B., Heck, D. E., Heindel, N. D., Young, S. C., Sinko, P. J., Casillas, R. P., Laskin, J. D., Laskin, D. L., & Gerecke, D. R. (2014). Therapeutic potential of a non-steroidal bifunctional anti-inflammatory and anti-cholinergic agent against skin injury induced by sulfur mustard. Toxicology and applied pharmacology, 280(2), 236–244. https://doi.org/10.1016/j.taap.2014.07.016

7. Function Pintigny, D., & Dachary-Prigent, J. (1992). Aprotinin can inhibit the proteolytic activity of thrombin. A fluorescence and an enzymatic study. European journal of biochemistry, 207(1), 89–95. https://doi.org/10.1111/j.1432-1033.1992.tb17024.x

8. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755-763.

9. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol. 1994;235(5):1501-1531.

10. Hidden Markov Models and their Applications in Biological Sequence Analysis

11. The UniProt Consortium
UniProt: the Universal Protein Knowledgebase in 2023
Nucleic Acids Res. 51:D523–D531 (2023)

12. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.

13. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman Nucleic Acids Research (2020) doi: 10.1093/nar/gkaa913

14. Antonina Andreeva and others, SCOP2 prototype: a new approach to protein structure mining, Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D310–D314, https://doi.org/10.1093/nar/gkt1242

15. Antonina Andreeva and others, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D376–D382, https://doi.org/10.1093/nar/gkz1064

16. Knudsen M, Wiuf C. The CATH database. Hum Genomics. 2010 Feb;4(3):207-12. doi: 10.1186/1479-7364-4-3-207. PMID: 20368142; PMCID: PMC3525972.

17. Protein structure comparison service PDBeFold at European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/ssm), authored by E. Krissinel and K. Henrick: E. Krissinel and K. Henrick (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. D60, 2256---2268

18. # hmmscan :: search sequence(s) against a profile database
HMMER 3.3.2 (Nov 2020); http://hmmer.org/
Copyright (C) 2020 Howard Hughes Medical Institute.
Freely distributed under the BSD open source license.

19. Madden T, Camacho C. BLAST+ features. 2008 Jun 23 [Updated 2021 Mar 14]. In: BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK569839/

20. Typhaine Paysan-Lafosse and others, InterPro in 2022, *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D418–D427, https://doi.org/10.1093/nar/gkac993

21. Arun Prasad Pandurangan and others, The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D490–D494, https://doi.org/10.1093/nar/gky1130

22. Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology, 313(4), 903–919. https://doi.org/10.1006/jmbi.2001.5080

23. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I.
New and continuing developments at PROSITE
Nucleic Acids Res. 2012; doi: 10.1093/nar/gks1067
PubMed:23161676 [Full text] [PDF version]