Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Beatrice Pastorino[1]

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum Università di Bologna
Address Via Francesco Selmi, 3, Bologna, Italy

29.05.2023

1 Abstract

Motivation: Kunitz-type domain protease inhibitors are a ubiquitous type of protease inhibitor that play important roles in regulating protease activity in various natural systems and specifically the pancreatic Kunitz inhibitor, also known as bovine basic pancreatic trypsin inhibitor (BPTI), is one of the most extensively studied globular proteins. Understanding the molecular basis of these proteins can lead to the development of new drugs for treating diseases related to abnormal protease activity. The aim of this study is to build a model for the Kunitz domain using available structural information and use this model to annotate Kunitz domains in SwissProt.

Results: The profile hidden Markov model achieved to classify the Kunitz-type proteins at a best threshold of 10^{-2} with an accuracy of 0.999 and a Matthews correlation coefficient of 0.998. Out of 569126 negatives one protein was classified at this threshold as false positive and analyzing its structure on Uniprot it was found as a true positive that could have escaped the filtering during the collection of the negative dataset.

Contact: beatrice.pastorino@studio.unibo.it

Supplementary information: Supplementary material is available on *github*.

2. Introduction

Kunitz-type domain protease inhibitors are an important type of protease inhibitor and belong to the I2 family of protease inhibitors. Kunitz-type domains are ubiquitously found in natural systems as serine protease[1].The *bovine pancreatic trypsin inhibitor* (BPTI) is one of the most extensively studied globular proteins. It has proved to be a particularly attractive and powerful tool for studying protein conformation as well as molecular bases of protein/protein interaction(s) and (macro)molecular recognition. BPTI has a relatively broad specificity, inhibiting trypsin- as well as chymotrypsin- and elastase-like serine (pro)enzymes endowed with very different primary specificity[2].

Kunitz domains have been used as a framework for the development of new pharmaceutical drugs, like aprotinin, that was used as a medication to reduce bleeding during complex surgery. The drug aprotinin (Trasylol, previously Bayer and now Nordic Group pharmaceuticals), is a small protein bovine pancreatic trypsin inhibitor (BPTI). It is a monomeric (single-chain) globular polypeptide derived from bovine lung tissue. It has a molecular weight of 6512 and consists of 16 different amino acid types arranged in a chain 58 residues long[4][5] that folds into a stable, compact tertiary structure of the 'small SS-rich" type, containing 3 disulfides, a twisted β-hairpin and a C-terminal α-helix[6].The high stability of the molecule is due to the 3 disulfide bonds linking the 6 cysteine members of the chain .The basic structure can be represented by the following simplified illustration (see Figure 2).

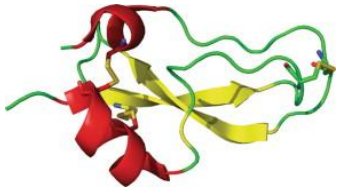

Figure 1 Structure of BPTI: In this structure (PDB ID: 1BPI) [3], a-helices are shown in red, b-sheets in yellow, and loop regions in green. Three conserved disulfide bonds are shown as sticks.

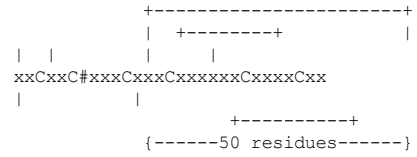

Figure 2 Simplified representation of the bovine pancreatic trypsin inhibitor domain. "C" denotes conserved cysteines that are involved in disulfide bonds. "#" denotes active site residue. Image: [8]

In this project Machine learning approaches have been used since they are widely adopted in the bioinformatic context [9]. Conservation patterns of a given protein domain may be represented by Markov chains. They describe discrete-time stochastic processes and are the core of *hidden Markov models* (HMM). They are used in the modelling of the likelihood of a residue to be present in a given position of a protein sequence. HMMs can be trained to identify a variety of patterns, which is necessary for domain

profiling. Protein sequences of a certain family share the same function which is reflected in their structure and their evolutionary relationship. Some positions within a protein family are more conserved than others and are typically the ones crucial for a functional active site.

These patterns can be exploited by a profile HMM. The topology of an HMM can be described as illustrated in Figure 3 [10][11]. The advantages of this approach are that it allows to characterize an entire protein family. Further it is based on formal statistics and probability and it allows to make libraries of hundreds of profile HMMs applicable to large scale data. The aim was to make use of the HMM to create a method for identifying the bovine pancreatic trypsin inhibitor domain.

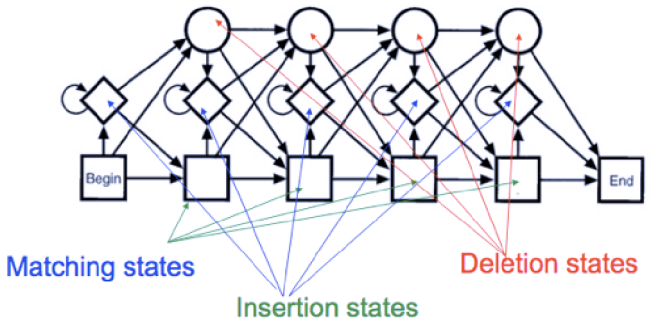

Figure 3 State diagram of an HMM for a sequence. The number of matching sates is given by the average sequence length in the family and is shown in blue. Insertion states are highlighted in green while deletion states are highlighted in red. Image modified from [11].

2. Methods

The workflow of this projects consists in six steps (see Figure 4)

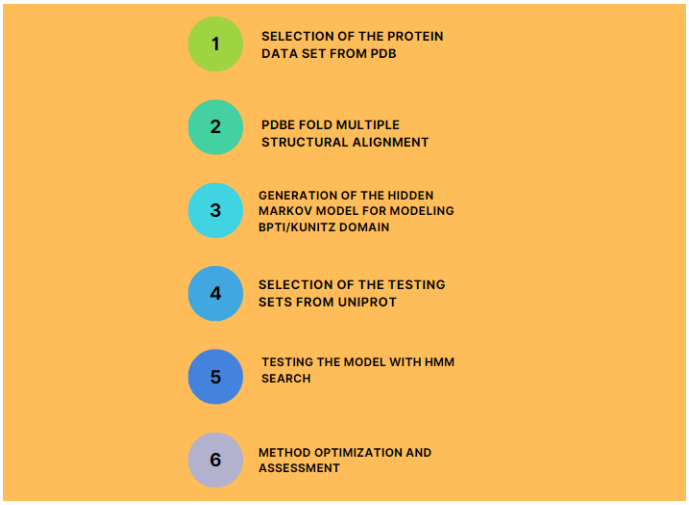

Figure 4 Schematized steps of the workflow used in this project.

Selection of the protein data set from PDB

The data set of proteins containing the Kunitz domain was retrieved with an advanced search among the structures on *Protein Data Bank* (PDB) [12,13], applying the following filters:

-PFAM domain: we included only proteins that match with the PFAM ID PF00014, associated with the Kunitz/Bovine pancreatic trypsin inhibitor (BPTI) domain.

-Length of polymer chain: we included only proteins with a length in the range 50-80 aa. to include merely the domains themselves.

-Resolution: we chose structures with a resolution max 3.0 Å. Structures with higher resolution would compromise the reliability of our model. This threshold of 3 Å is chosen to ensure appropriate modeling of interacting bonds and the α-carbons.

-sequence identity: 95%

QUERY: (Identifier = "PF00014" AND Annotation Type = "Pfam") AND Polymer Entity Sequence Length = [49 - 80] AND Refinement Resolution <= 3. After this filtering process there were 33 structures. The Custom Tabular Report was created with the following options selected: PDB ID, AUTH ASYM ID. After downloading this file from PDB, it was cleaned to achieve for every PDB ID the corresponding chain in a format perfect to upload on PDBe Fold.

PDBe Fold multiple sequence alignment

PDBe Fold v2.59. (src3) [14–17] is an algorithm that uses elements of secondary structure (SSE) to do multiple structural alignment. An alternative to PDBe Fold is mTM-align. In the results from the session on PDBe fold, in the matrix containing all the RMSD (between all against all), the majority of the RMSD are below 1 Å , which means they are quite good alignments.The file of the resulting MSA was downloaded in FASTA format and cleaned so that the file is in a sort of clustal format.

Generation of the Hidden Markov Model for modeling BPTI/Kunitz domain

For generating the profile HMM we employed HMMER v3.3.2 [19]. It is a program designed to identify distant homologs while depending on the strength of its underlying probability models. The program generates the profile with the help of a specific scoring system for deletions, substitutions and insertions based on the transition probability. To account for uncertainty HMMER3 calculates match scores by considering all possible alignments weighted by their relative likelihood [18][19]. We used this profile HMM to generate a sequence logo on the Skylign Web application illustrating the conserved residues (see Figure 5) [20][21].By studying the seq logo it's clear that there are 6 cysteins(C) conserved. They are important for keeping the structure compact with three disulfide bonds. Then there are other sites that are potentially important for the function of the protein such as the tyrosines (Y).

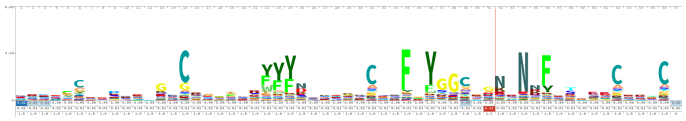

Figure 5 Sequence logo of the profile HMM of the BPTI domain. Image was generated using Skylign.org

Selection of the testing sets from Uniprot

For the testing of our profile HMM we had to use a dataset capable of validating the classification. The logic choice was using data from SwissProt, a part of UniprotKB containing only manually annotated proteins [22][23]. We retrieved protein sequences containing the BPTI domain for the positive set and proteins lacking such domain for the negative set. The query for the positive set was: query: (xref:pfam-pf00014) AND (reviewed:true) and returned 384 positive hits. The 569126 negative hits were retrieved with thequery: NOT (xref:pfam-pf00014) AND (reviewed:true). Both sets were saved locally, the negative dataset as FASTA file and the positive dataset as list file because it is important to note that for training and validation the positive

set must exclude the sequences used in the training of the profile HMM. So The main idea is that the set of ids used for building the model should be "mapping ids" option on Uniprot; then, after the removal of these ids, the next step consisted of doing again the mapping ids on Uniprot with the left uniprot ids to get the FASTA file.

Testing the model with *HMM search*

hmm search [19] takes as input an hmm file and a fasta sequence input.
To facilitate file parsing and help the program to work in a more robust manner the following options were used:

--noali : don't output alignments, so output is smaller, because we only want the E-value for classification.

--max : Turn all heuristic filters off (less speed, more power)

With hmm search a set of sequences was matched with the HMM model and the resulting file was cleaned.The cleaned file has 368 sequences like the file of the positive test dataset before running hmm search. So it means every sequence was found as a match, all these proteins of the positive test dataset contain the kunitz domain as expected.

The same procedure was done with a negative test dataset (without kunitz domain): first hmm search and then the resulting file was cleaned.The resulting file has 33 sequences. So out of 569126 sequences that were provided in input the result is 33 matches. It was necessary to extract the identifier and the e-value and finally to add "0". The same step was done with the positive file with kunitz matches but "1" was added instead of "0" .

the file with 33 proteins has a lot less proteins than the 569126 in the list of negatives at the beginning, so it was necessary to recover the ones that were missing; to have the same number of sequences as in the negative dataset at the beginning (569126), the number of missing lines can be added with an e-value that is higher than the highest e-value. So after checking in the cleaned file from hmm search with negative data set that the highest e-value is 26, to the ones that have not been identified with the hmmsearch an e-value of 100 was assigned. The missing lines were added checking what the missing identifiers are and assigning them higher e-value and the class 0.

Method optimization and assessment

2-fold-cross-validation: shuffle of these 2 files and selection of half of the set in a file and half of the other set in another file to make them comparable and reversible. Creation of a python program that takes in input a file containing protein ID, e-value and protein class (0/1) to calculate the performance of the classification method at different e-value thresholds. After running this python program on the file set 1 and file set 2 and the output was the threshold in a range of values, the accuracy(acc), Matthew correlation coefficient (mcc) and the confusion matrix(cm). It computes a confusion matrix which contains true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) (see Table 1). The program computed in addition true positive rate (TPR), false positive rate (FPR).

		Actual class	
		Positives	Negatives
pre-dicted class	Positives	TP	FP
	Negatives	FN	TN

Table 1 Confusion matrix showing the true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN)

Optimization: prediction could be optimized by changing the E-value threshold. So once it was found the good threshold of e-value to be 0.01 it can be used as cutoff for running hmm search again with the negative test set to see if the number of matches changes. After cleaning the resulting file, the same procedure was done with the positive dataset.
ROC: The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. A python program called ROC.py was created and applied to the two sets set 1 and set 2 removed from the positive dataset list from uniprot. To do this we had to convert the list of pdb ids used for the training into uniprot ids with the and after extracting from the output file the values of FPR and TPR with different thresholds, importing matplotlib it was created a plot of the ROC.

2. Results

Positive dataset: 384 proteins containing Kuntiz domain were found and then after the removal of the 33 ids used for building the model a file that represents the positive dataset was created, it contains 368 sequences.
Negative dataset: 569126 proteins not containing Kunitz domain were found and that will represent the negative test and it was downloaded as FASTA file.

hmm search results
the output of step 5.2 (see supplementary material, materials_methods_steps) is a cleaned file of kunitz domain proteins deriving from hmm search (see supplementary material, kunitz_clean.out) that has 368 sequences like the file of the positive test dataset (see supplementary material, kunitz_clean.fasta) before running hmm search. So it means every sequence was found as a match, all these proteins of the positive test dataset contain the kunitz domain as expected.
The output of step 5.3 (see supplementary material, materials_methods_steps) is a cleaned file of non kunitz domain proteins deriving from hmm search (see supplementary material, nonkunitz_new.out) that has 33 sequences. So out of 569126 sequences that were provided in input the result is 33 matches.
2-fold-cross-validation: from step 6.2 (see supplementary material, materials_methods_steps file). the outputs of the performance of the model are the following:

Best Results Were Achieved at an E-value of 10^{-2}

Set	Thr.	ACC	MCC
test 1	10^{-2} -10^{-6}	1.000	1.000
test 2	10^{-2}	0.999	0.997
test 1+ 2	10^{-2}	0.999	0.998

Table 3 Statistics of testing (set test 1, set test 2, set test 1 and 2).

Optimization results: in step 6.3 (see supplementary material, materials_methods_steps file) one of the outputs is a file (see supplementary material, non_kunitz_new2.search) deriving from the optimization with the best threshold for running hmm search. After the cleaning, the result is a file (see supplementary material, non_kunitz2.out) that has less sequences (1) than the ones (33) found before in the cleaned file called nonkunitz_new.out (see supplementary material, nonkunitz_new.out) step 5.4 (see supplementary material, materials_methods_steps) so it means with a better threshold than the default one hmm search found less matches. all the sequences in the negative dataset should not be recognised by hmm search as matches but actually 1 protein passed the e-value threshold. There are some possible reasons why this happened:
This protein could be false positive because the model might have detected some regions in the negative sequences that resemble the kunitz domain, but are not actually kunitz domains. This can happen if the model is not specific enough, or if the sequence contains regions with similar sequence patterns to the kunitz domain.
Incomplete filtering: It is possible that one of the negative sequences used for testing still contains residues that are part of the kunitz domain, but were not identified by the filtering method. These residues could be responsible for the detection of a false positive hit.
P84555 · TIQ7_RHISA protein is a serine protease inhibitor, on
Uniprot it has a 2/5 annotation score and seems to contain kunitz domain(see figure 6)

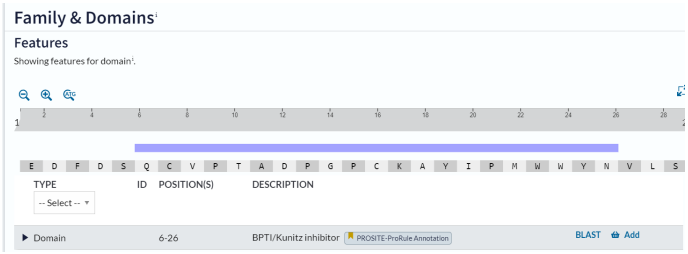

Figure 6 Uniprot [22][23] family and domains, protein P84555. The kunitz domain is present in position 6-26.

so the hmm model built was correct on identifying it as kunitz domain protein, it's a true positive. The problem could have been on the negative testing set phase where the filtering from uniprot went wrong on this protein.

ROC results: the area under the ROC curve [25] indicates the performance of the method that is threshold-independent (because all the possible thresholds were already calculated) while other performance methods that were done previously are threshold dependent. ROC is dependent only on the predictor/model.
The closer to 1, the better the performance (graphically means that minimum level of false positive and maximum level of true positive); closer to diagonal (area of a triangle) the area will be 0,5 so a random predictor. The ROC curve obtained for the hidden markov model built in this project was plotted (see Figure 7) after running a python code called ROC.py on the sets computing two lists with true positive rate (TPR), false positive rate (FPR).

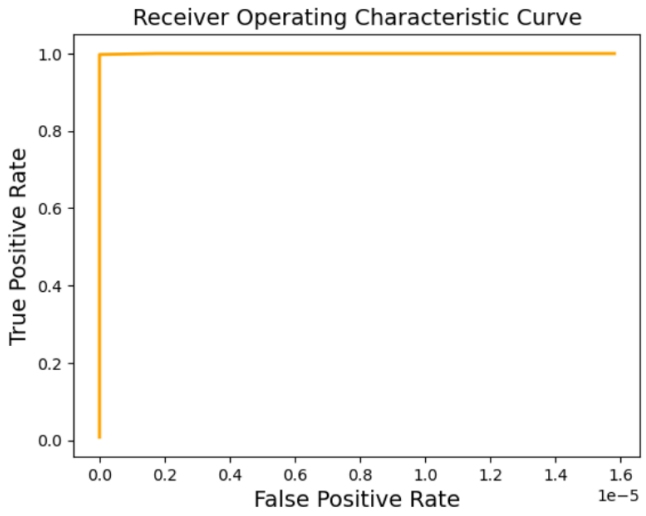

Figure 7 In orange the receiver operating characteristic curve obtained by our plotting our false positive rate (x - axis) against our true positive rate (y - axis).

2 Conclusion

In conclusion, the two specific aims of this study have been successfully achieved. Firstly, a model for the Kunitz domain has been built using available structural information. This model provides valuable insights into the structural and functional characteristics of this important domain, and has the potential to advance our understanding of Kunitz domain-containing proteins. hmm search with the optimal e-value found only 1 match from the negative dataset. After analyzing the sequence on SwissProt it seems that this protein really contains the BPTI/kunitz domain (in position 6-26). This could mean that it is a true positive and the negative dataset was collected with a bad filtering on this protein.Overall, the model built is a good model for BPTI/kunitz domain as proved by the ROC curve. Finally this study has demonstrated the value of developing accurate models for protein domains, and the potential of these models for advancing our understanding of protein structure and function.

3 Acknowledgements

Thanks to all professors who took their time to answer questions and to my colleagues who were available for discussions.

4 Funding

project self financed.

Conflict of Interest: none declared.

5 References

1. Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem J. 2004;378(Pt 3):705-16. doi: 10.1042/BJ20031825
2. Ascenzi P, Bocedi A, Bolognesi M, Spallarossa A, Coletta M, De Cristofaro R, et al. The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. Curr Protein Pept Sci. 2003;4(3):231-51. doi: 10.2174/1389203033487216
3. Parkin S, Rupp B, Hope H. Structure of a Kunitz-type inhibitor of human tissue kallikrein, hTKI. Nat Struct Biol. 1996;3(5):365-9. doi: 10.1038/nsb0596-365.
4. Mannucci PM (July 1998). "Hemostatic drugs". The New England Journal of Medicine. 339 (4): 245–53. doi:10.1056/NEJM199807233390407
5. Mahdy AM, Webster NR (December 2004). "Perioperative systemic haemostatic agents". British Journal of Anaesthesia. 93 (6): 842–58. doi:10.1093/bja/aeh227
6. Richardson JS (1981). "The anatomy and taxonomy of protein structure". Advances in Protein Chemistry Volume 34. Advances in Protein Chemistry. Vol. 34. pp. 167–339. doi:10.1016/S0065-3233(08)60520-3
7. Nixon A, Wood CR. Engineered protein inhibitors of proteases. Current Opinion in Drug Discovery & Development. 2006;9(2):261–8.
8. Summary: Kunitz/Bovine pancreatic trypsin inhibitor domain [Internet]. [cited 2020 May 10]. Available from: http://pfam.xfam.org/family/PF00014
9. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min [Internet]. 2017 Dec 8 [cited 2023 May];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660/

10. Momand J, McCurdy A. Concepts in bioinformatics and genomics. International ed. New York: Oxford University Press; 2017. xvi+448.

11. HIDDEN MARKOV MODELS IN MULTIPLE ALIGNMENT. 2 HMM Architecture Markov Chains What is a Hidden Markov Model(HMM)? Components of HMM Problems of HMMs. - ppt download [Internet]. [cited 2023 May]. Available from: https://slideplayer.com/slide/4970773/

12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

13. Bank RPD. RCSB PDB [Internet]. [cited 2023 April]. Available from: https://www.rcsb.org/

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct 5;215(3):403–10.

15. Krissinel E, Henrick K. Multiple Alignment of Protein Structures in Three Dimensions. In: R. Berthold M, Glen RC, Diederichs K, Kohlbacher O, Fischer I, editors. Computational Life Sciences [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005 [cited 2023 May]. p. 67–78. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 3695). Available from: http://link.springer.com/10.1007/11560500_7

16. PDBefold.pdf [Internet]. [cited 2023 April]. Available from: https://www.ebi.ac.uk/pdbe/docs/Tutorials/workshop_tutorials/PDBefold.pdf

17. PDBe < Fold < EMBL-EBI [Internet]. [cited 2023 April]. Available from: https://www.ebi.ac.uk/msd-srv/ssm/

18. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013 Jul 1;41(12):e121–e121.

19. HMMER [Internet]. [cited 2023 April]. Available from: http://hmmer.org/

20. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics. 2014 Jan 13;15(1):7.

21. Skylign [Internet]. [cited 2023 April]. Available from: http://skylign.org/

22. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

23. UniProt [Internet]. [cited 2023 April]. Available from: https://www.uniprot.org/

24. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014 Nov 15;30(22):3276–8.

25. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.