Laboratory of Bioinformatics 1 Module B

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

Becchi Tommaso

University of Bologna, Master Degree Course in Bioinformatics

Abstract

Motivation: This work aims to build a profile for the Kunitz-type protease inhibitor domain using a Hidden Markov Model-based approach. This method starts with a multiple alignment between structures that contain this domain. Once you get the model, you need to test it comparing its performances with a positive set and a negative one.
BPTI (Bovine Pancreatic Trypsin Inhibitor) is the active domains of proteins that inhibit the function of proteases and are called Kunitz-type protease inhibitors [1]. Their structure contains an alpha+beta fold with many cysteines that stabilize the structure forming three disulfide bridges.
Results: The obtained HMM model is a good predictor for the classification of Kunitz-type proteins. The confusion matrices obtained from different Training and Testing sets have an Accuracy of approximately 100% and a Matthews Correlation Coefficient's values very close to 1. This indicates a reliable prediction for most of cases.
Furthermore, it makes possible to investigate SwissProt annotation to find some issues which better explain our results.

1 Introduction

The Kunitz-type domains are the active domains of proteins that inhibit the function of protein degrading enzymes. They are relatively small: the length of this domain is about 50 to 60 aminoacids and their molecular weight is 6 kDa[2].
Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP) [3], and tissue factor pathway inhibitor (TFPI) [4].
BPTI is a monomeric single-chain globular polypeptide derived from bovine lung tissue. It is composed of 58 residues that allow obtaining a stable tertiary structure, containing 3 disulfides bonds, a twisted β-hairpin, and a C-terminal α-helix (Figure 2) (Figure 3) [1]. The basic structure of such a type of inhibitor is shown in the following schematic representation:

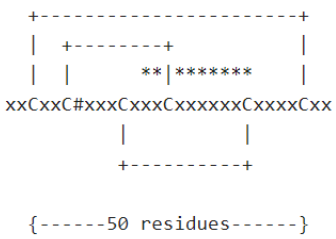

Fig 1. Schematic representation of BPTI-Kunitz domain superfamily (InterPro). This scheme represents the main elements of this domain in InterPro. 'C' are the conserved cysteines involved in a disulfide bond. '#' is the active site residue and '*' are the position of the pattern

Fig. 2. Schematic representation of the PBTI-Kunitz domain (PDB). This image is obtained from the PDB page of the structure 5YV7. It is possible to observe the main elements of this domain: 3 disulfide bonds, 2 β-strands, and one terminal α-helix.

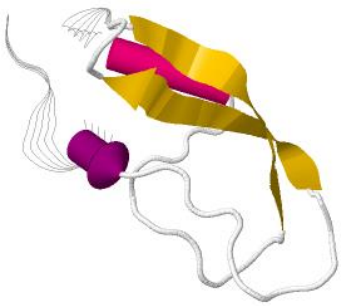

Fig. 3. Representation of the BPTI-Kunitz domain structure. This image is obtained from the PDB page of the structure 5YV7 and it allows us to observe the main elements. 2 β-strands are colored in yellow and the terminal α-helix is colored in purple.

This domain has high stability because of its 3 disulfide bonds linking the 6 cysteine members of the chain.
In figure 1 it is possible to observe the presence of a residue inside the domain which has the function of active site: it is a basic lysine on the exposed loop. It binds very tightly in the specificity pocket at the active site of trypsin and inhibits its enzymatic action [5].

Thanks to this conserved lysine the proteins with this domain can perform the cleavage of the trypsinogen precursor during storage in the pancreas when small amounts of the major digestive enzyme trypsin are produced. Both the Lys/Arg15 and Cys residues (at positions 5, 14, 30, 38, 51, and 55) are highly conserved and allow to identify this domain.
In UniProt 359 reviewed sequences are annotated with the Pfam code for this domain (PF00014) according to the rule PROSITE-ProRule: PRU00031[6].

Starting from available structural information, it is possible to develop a Hidden Markov Model of the domain and test it [7][8]. Hidden Markov Models (HMMs) are statistical models that use parameters estimated from a training set of aligned sequences.
After the HMM is built, it can be used to search the database for other sequences that are members of the given protein family or contain the given domain. The HMMs are very accurate to distinguish sequences that are members of these families from non-members [9].

2 Material and Methods

Databases

The structures used to create the HMM of the domain are selected from PDB (release 2020_04) [10]. The selected sequences have these properties: PF00014 as Pfam annotation [11], resolution less than or equal to 3.5 Å, and the number of polymer residues less than 100. This query identifies 39 structures.

Datasets

Both the negative set and the positive set are obtained from UniProt (release 2020_04) [12]. The positive set contains all the reviewed sequences annotated with Pfam code PF00014, while the negative one contains all the reviewed sequences not annotated with Pfam code PF00014. There are 359 sequences in the positive set and 561894 sequences in the negative set.

Computational methods

PDBeFold (v2,59) [13] is an online tool implemented in PDBe which allows performing both pairwise and multiple comparisons and 3D alignment of protein structures.
HMMER (v3.3) [14] is used to perform all the analyses related to the HMM. This tool includes many programs which allow performing different operations:

- *hmmbuild* program reads a multiple sequence alignment file and builds a new profile HMM.
- *hmmsearch* reads an HMM file and searches for significantly similar sequence matches in a set of sequences.

Skylign (https://skylign.org/) [15] allows us to obtain the HMM logo giving in input the file with the parameters of the model.
BLAST (v2.2.26) [16] is a software distributed by NCBI which finds regions of similarity between biological sequences. It compares nucleotide sequences to a sequence database and calculates the statistical significance:

- *blastclust* automatically clusters DNA sequences based on pairwise matches found using the BLAST algorithm.
- *formatdb* is used to index protein or nucleotide databases so that they are formatted for searching.
- *blastpgp* performs gapped blastp searches and can also be used to perform iterative searches in psi-blast and phi-blast mode.

Generating the HMM of the domain

The 39 sequences derived from the structures selected on PDB are clustered together with BLUSTClust to eliminate the redundancy of the information. The following parameters are used: similarity threshold equal to 99% and minimum length coverage also equal to 99%.
The program generates 14 clusters and one entity for each cluster is used to perform a multiple structural alignment on PDBeFold. The result is a table with all the RMSD values for each of the pairs among all possible pairs. Analyzing this table, it is evident that the second cluster is very different from the others: it has an RMSD greater than 3 Å with each one of the other sequences.
Checking the possible cause on PDB, it seems that the structures in this cluster have the BPTI domain, but they also have other domains that lead to this difference.
This cluster is eliminated and a new alignment on PDBeFold is performed. Observing the resulting table, it is possible to observe that now all the clusters are similar so the multiple alignment can be used to calculate the HMM.
Using Skylign it is possible to obtain the logo of the HMM (Figure 4). The most visible aspect is the conservation of the 6 cysteines involved in the disulfide bonds.

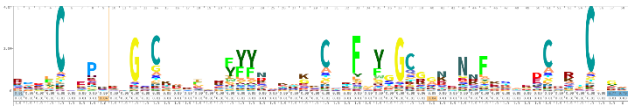

Fig. 4. Logo of the HMM obtained on Skylign [17]

Testing the method

From UniProt it is possible to obtain the positive and the negative sets. Positive set contains 359 sequences but some of them are the ones used to develop the HMM, so they are no informative during the test. They must be eliminated and to identify them BLAST is used. The sequences in the positive set which have 100% similarity with the sequences in the set used for the generation of HMM are eliminated. In such a way the positive set contains 354 sequences.
To perform the two-fold cross-validation of the model, both sets are split into half (2 sets with 177 sequences for the positives and 2 sets with 280947 sequences for the negatives) so that we can exploit the first half and the second one respectively for the *Training* and the *Testing* set.
Each one of these 4 sets is compared to the HMM with hmmsearch to obtain statistical values for the similarity among the sequences and the model. In this comparison the *Z* parameter is set equal to 1. This parameter allows us to eliminate the problem of the relevant numerical difference between positive and negative sets. The *max* parameter is also activated: it turns off all filters and runs full forward/backward postprocessing on every target to increases sensitivity.
In the results it is possible to observe the E-value for the similarity of each sequence with the model both at the level of the entire sequence and the level of the best domain. Among these two results the most significant is the one which statistically evaluates the similarity of the best domain because it is the element of the protein which is biologically relevant.
Knowing the E-value for the sequences in both the positive and negative sets, it is possible to calculate the confusion matrix. To obtain this matrix is necessary to choose a threshold that distinguishes between positive and negative results (Table 1).

	POSITIVE SET	NEGATIVE SET
POSITIVE RESULT	True positive	False positive
NEGATIVE RESULT	False negative	True negative

Table 1. Confusion matrix. This is a schematic representation of a confusion matrix. For a given threshold it is possible to identify four possible results. True positives are those sequences which belong to the positive set and whose E-value is less than the threshold. False negatives are those sequences which belong to the positive set and whose E-value is greater than the threshold. False positives are those sequences which belong to the negative set and whose E-value is less than the threshold. True negatives are those sequences which belong to the negative set and whose E-value is greater than the threshold

Measuring the performance

Using a python script, it is possible to automatically fill the confusion matrix for a large set of thresholds to find which is the best. The performance [18] of each matrix is evaluated with 2 parameters: accuracy (ACC) and Matthew's correlation coefficient (MCC).

Accuracy is the ratio of correct predictions to total predictions made.

$ACC = (TP + TN)/(TP + FN + TN + FP)$

MCC [19] is a correlation coefficient with a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

$MCC=[(TP*TN)-(FP*FN)]/\sqrt{[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]}$

3 Results

Calculating the accuracy and the MCC for the *Training* set with different thresholds it is possible observing that the model performs very well in distinguishing the sequences which have this domain to those which do not have it (Table 2).

TH	ACC	MCC	FP	FN
1	0,716167954	0,039808794	79792	0
0,1	0,963400492	0,127642265	10289	0
0,01	0,995966193	0,366696725	1134	0
0,001	0,999783014	0,862285339	61	0
1E-04	0,999953757	0,965161044	13	0
1E-05	0,999989329	0,991626358	3	0
1E-06	0,999992886	0,994394176	2	0
1E-07	0,999996443	0,99718528	1	0
1E-08	0,999996443	0,997169365	0	1
1E-09	0,999989329	0,991483913	0	3
1E-10	0,999985771	0,988628957	0	4
1E-11	0,999985771	0,988628957	0	4
1E-12	0,999982214	0,985765752	0	5
1E-13	0,999971543	0,977125925	0	8
1E-14	0,999953757	0,962554311	0	13
1E-15	0,999943086	0,953704734	0	16
1E-16	0,999921743	0,935755121	0	22
1E-17	0,999864828	0,886117673	0	38
1E-18	0,999822143	0,846986689	0	50
1E-19	0,9998008	0,826728241	0	56

Table 2. Performances of the HMM with different thresholds tested on the *Training* set. The highlighted rows are those with better performance. The number of identified false positives (FP) and false negatives (FN) for each threshold are also reported.

Accuracy and MCC are very close to 1 for each threshold in the selected range (except for very permissive thresholds in the 0,01-1 range). With 1E-07 and 1E-08 as threshold, we have the best performances.

1E-07 identifies only one false positive and 0 false negatives while 1E-08 identifies only one false negative and 0 false positives. It means that only one sequence out of 281124 (177 in the positive set and 280947 in the negative one) is assigned to the incorrect set. It seems that these thresholds are very good to evaluate if a sequence has or not the domain.

Nevertheless, it is necessary to evaluate a further estimation of the correctness of the predictor.

The same analysis is performed on the *Testing* set. The results for the thresholds previously chosen and those with a close value are summarized in Table 3.

TH	ACC	MCC	FP	FN
1E-06	0.99996785	0.975009038	7	2
1E-07	0.999982214	0.985910511	3	2
1E-08	**0.999989328**	**0.991500060**	**1**	**2**
1E-09	**0.999989328**	**0.991500060**	**1**	**2**
1E-10	**0.999989328**	**0.991500060**	**1**	**2**
1E-11	**0.999989328**	**0.991500060**	**1**	**2**
1E-12	**0.999989328**	**0.991500060**	**1**	**2**
1E-13	**0.999989328**	**0.991500060**	**1**	**2**
1E-14	0,999982214	0,985782139	1	4

Table 3. Performances of the HMM with different thresholds tested on the *Testing* set. The results for the thresholds previously chosen are highlighted.

It is possible to observe that the model has a good performance also for this set. The set of thresholds that give us the best results is slightly different. The thresholds in the range from 1E-08 to 1E-13 give us the same results in terms of performance. 1E-08 is the value which belongs to the best results for both the sets. Then it is possible to choose 1E-08 as the best threshold to distinguish between structures that have this domain and structures which do not have it.

With this threshold 1 false positive and 2 false negatives are identified in this set.

Further validation is obtained by calculating the ROC curve [20]. It consists of a graph in which the true positive rate (TPR) is plotted with the false positive rate (FPR) as the threshold changes. The higher is the area under this curve the better is the model in terms of precision and accuracy.

Watching the resulting graphs (Figure 5) is evident that this model performs very well because the area under the curve is the maximum possible.

Observing the results for the selected best threshold it is important to focus the attention on the identified false positives and false negative. It is necessary to investigate them to understand which can be the possible issues. Setting the threshold to 1E-08, there is 1 false negative in the *Training* set and 1 false positive plus 2 false negatives in the *Testing* set. It is possible to know the UniProt Id of these sequences and the results are summarized in Table 4.

	False positive	False negative
Training set		Q11101
Testing set	G3LH89	D3GGZ8 O62247

Table 4. UniProt Ids of the identified false positives and false negative with 1E-08 as a threshold.

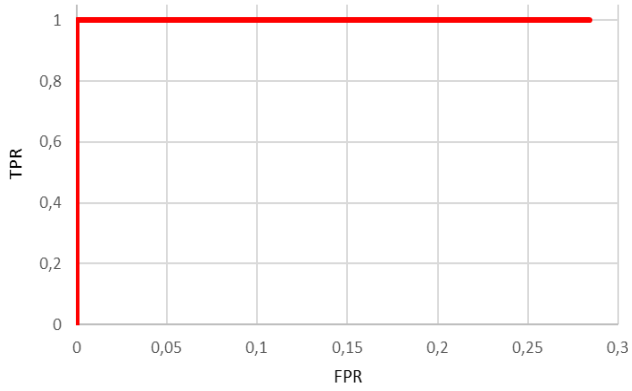

Fig. 5. ROC curve of the generated model. This graph represents how FPR and TPR change at a different threshold in the confusion matrix obtained with the comparison between the model and the *Training* set. The resulting graph for the *Testing* set is very similar so it is possible to assert that the model has a good performance.

FPR is calculated as the ratio between FP and FP+TN.

TPR is calculated as the ratio between TP and TP+FN.

Analyzing the information on UniProt for these Ids it is possible to make some hypotheses on the possible issues.

G3LH89 is a false positive so it belongs to the negative set, but this Id identifies a protein called *Kunitz-type serine protease inhibitor Bi-KTI* (Organism: *Bombus ignitus*). In "Family and Domain" annotation this sequence is annotated with the BPTI/Kunitz inhibit domain, but it is not annotated with Pfam code PF00014. So, it is possible to suppose that this sequence contains the domain, but it is not correctly annotated on Pfam.

False negatives are sequences annotated with Pfam code PF00014 which have low similarity with the model of the domain.

D3GGZ8 identifies *Kunitz-type protein bli-5* (Organism*: Haemonchus contortus*). This sequence has a BPTI domain with a length of 71 amino-acids while our domain normally has a length in the range of 50-60 aminoacids.

Q11101 and O62247 entries do not have any feature which can cause their difference with the model. What they have in common is that both these Ids identify sequences that belong to the organism *Caenorhabditis Elegans*. It is possible to suppose the cause of their low similarity is the phylogenetic distance of this organism compared to the organisms to which the structures used to build the model belong.

4 Conclusions

It is possible to assert that the model generated with the HMM method describes properly the domain. The confusion matrices resulting from a comparison between the model and positive and negative sets have a good performance.

The results obtained with the *Training* and the *Testing* sets are quite similar. They give us a range of thresholds with which there are very few errors in the discrimination between positives and negatives.

It means that that the model is useful to distinguish between sequences that have the domain and sequences which do not have it. Using 1E-08 as a threshold for the discrimination of the E-values, there are only 4 sequences out of 562248 (354 in the positive set plus 561894 in the negative one) that are incorrectly assigned. The causes are supposed to be the presence of a slightly different form of the domain, the derivation from an organism that is phylogenetically distant to the others, and an incorrect annotation on UniProt.

Supplementary

In the Supplementary materials the Python scripts used in the procedure and the statistical results of multiple alignment on PDBeFold are reported.

References

1. Durani, V. & Magliery, T. J. Protein Engineering and Stabilization from Sequence Statistics. in Methods in Enzymology vol. 523 237–256 (Elsevier, 2013).
2. Ley, A. C. (54) KUNITZ DOMAIN PEPTIDES (75) Inventors: Robert Charles Ladner, Ijamsville. 120.
3. O'Brien, R. J. & Wong, P. C. Amyloid Precursor Protein Processing and Alzheimer's Disease. Annu. Rev. Neurosci. 34, 185–204 (2011).
4. Petersen, L. C. et al. Inhibitory Properties of a Novel Human Kunitz-Type Protease Inhibitor Homologous to Tissue Factor Pathway Inhibitor †. Biochemistry 35, 266–272 (1996).
5. Nixon AE, Wood CR. Engineered protein inhibitors of proteases. Current Opinion in Drug Discovery & Development. 2006 Mar;9(2):261-268.
6. Hulo, N. The PROSITE database. Nucleic Acids Research 34, D227–D230 (2006).
7. Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755–763 (1998).
8. Eddy, S. R. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. PLoS Comput Biol 4, e1000069 (2008).
9. Eddy, S. R. A New Generation Of Homology Search Tools Based On Probabilistic Inference. In Genome Informatics 2009 205–211 (Published By Imperial College Press And Distributed By World Scientific Publishing Co., 2009). Doi:10.1142/9781848165632_0019.
10. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.
11. El-Gebali, S. et al. The Pfam protein families database in 2019. Nucleic Acids Research 47, D427–D432 (2019).
12. The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46: 2699 (2018)
13. E. Krissinel and K. Henrick (2005). Multiple Alignment of Protein Structures in Three Dimensions. In: M.R. Berthold et.al. (Eds.): CompLife 2005, LNBI 3695, pp. 67--78. Springer-Verlag Berlin Heidelberg.
14. Eddy, S. R. HMMER User's Guide. 229.
15. Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics 15, 7 (2014).
16. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997).
17. Schuster-Böckler, B., Schultz, J. & Rahmann, S. HMM Logos for visualization of protein families. BMC Bioinformatics 8 (2004).
18. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437 (2009).
19. Boughorbel, S. & Jarray, Fethi & El-Anbari, Mohammed. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE. 12. e0177678. 10.1371/journal.pone.0177678.
20. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159 (1997).