Bioinformatics, 2021, 16–8

Laboratory of Bioinformatics 1b

Building a Hidden Markov Model to Annotate a Kunitz Domain Protease Inhibitor

Francesca Catalogne[*1]

[1]Department of Biotechnology and Pharmacy, University of Bologna, Italy.

*To whom correspondence should be addressed.

Abstract

Motivation: The Kunitz type protease inhibitory domain, also known as Bovine Pancreatic Trypsin Inhibitor (BPTI) is one of the most studied globular proteins due to its use in surgical and therapeutic interventions of reducing hemorrhagic complications. The aim of this work is to build and test a Hidden Markov Model (HMM) to classify new proteins that potentially contain the Kunitz domain in order to explore the potential for novel therapeutic intervention. The methods used in this project would permit the recognition of new BPTI-containing proteins.

Results: The HMM generated from the structural alignments have the potential to be a valuable predictor of Kunitz-type proteins, as there was sufficient accuracy of the Correlation Coefficients.
Contact: Francesca.catalogne@studio.unibo.it

Supplementary information: Supplementary data are attached

1 Introduction

Proteins containing the Kunitz active domains derive from a diverse family of proteins of 1-6 domains specifically known for their function of protease inhibition (Karim and Adamson, 2012). There are various types of Kunitz protease inhibitors such as Alzheimer's amyloid precursor protein, tissue factor pathway inhibitor and the classic bovine pancreatic trypsin inhibitor (BPTI) that was the first to be described, is widely studied and touted to be one of the smallest and simplest globular proteins studied as it inhibits the function of proteases, chymotrypsin and elastase like serine proenzymes with different specificity ("Kunitz Domain | ScienceDirect Topics," n.d.; Paolo et al., 2003). This inhibitor has been used as a drug called Aprotinin, which was used within cardiopulmonary surgeries, blood and liver transplants as an antihemorrhagic intervention but has since been withdrawn due to increased risk of complications and death (Dittrich and Kanchanawarin, n.d.). The Kunitz domains were first functionally resolved in ticks and have been widely studied since ("Kunitz Domain | ScienceDirect Topics," n.d.). A contributing hallmark factor to the stability of this BPTI protease inhibitor is the three disulfide bonds which are found in 6 cysteine residues (Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51) formed by oxidative folding ("Disulfide Bond | ScienceDirect Topics," n.d.). Due to its hallmark features of stability it is a useful model for studying protein conformation and the molecular bases of protein/protein interactions and molecular recognition (Paolo et al., 2003). The primary structure of the BPTI is made up of a single domain of 58 amino acid residues with a bond pattern of C1-C6, C2-C4, and C3-C5. It has a molecular weight of 6512 kDa and contains a twisted B-hairpin and C-terminal alpha helix. It contains 10 positively charged lysine and arginine side chains, and 4 negatively charged aspartates and glutamates, thus leading to a basic protein structure (Krokoszynska et al., 1998). The crystal structure of BPTI has been resolved using Nuclear Magnetic Resonance and X-ray crystallography, which has revealed its compact pear shape with a maximum dimension of about 30Å (Dittrich and Kanchanawarin, n.d.). The secondary structure reveals two α helical regions, H1 and H2, that form a one and a half turn 310-helix near the N-terminus and a three turn α-helix near the C-terminus. The central part of the protein contains two β-strands, B1 and B2, that are connected by a turn, T1, and which form an anti-parallel β- hairpin loop (Dittrich and Kanchanawarin, n.d.). Therefore, this structure is considered disulfide rich alpha and beta folds. This paper will transfer the structural knowledge from what is known to generate a Hidden Markov model from the PDB database ("RCSB PDB," n.d.), generating a multiple sequence structural alignment to build a Hidden Markov Model (HMM) as a method of binary classification (Eddy, 2011). Consequently, these performance results were tested against the proteins stored in Uniprot SwissProt database (The UniProt Consortium, 2019).

1 Methods

2.1 Dataset Selection, Structural Similarity Screening

Data regarding the protein was referenced by the RCSB Protein Data Bank (PDB) (“RCSB PDB,” n.d.) using the advanced search tool and the following criteria with ‘and’ Boolean keywords:

- Must contain the PF00014 Pfam ID domain that corresponds to BPTI/Kunitz
- Must contain a 3.5Å of less than or equal to resolution
- Must contain a sequence with 40-80 residues, upper included

The information retrieved 152 results, which was then downloaded into a .csv file with the following columns: entry ID, PDB ID, sequence, polymer entity sequence length, and auth asym ID. This dataset was manually analyzed for the removal of sequences that do not equate to the Kunitz domain of each protein and also reshaping was done to account for those residues between 40-80 in length to remove possibilities of co-crystalized structures i.e those with multiple domains. However, there is still the problem of redundancy that remained after this reshaping.

The dataset was then compared against the manually selected Kunitz type structure found in the PDB (3TGI; chain I) with a focus on only the chains with the Kunitz domain of interest. The first crystallized BPTI Kunitz domain, 3TGI complexes with wild-type rat anionic trypsin complexed with BPTI and its crystal structure has been resolved using X-ray diffraction with a resolution of 1.8Å which validates its quality. To search protein structures against the referenced protein and chain I, PDBeFold was utilized (“PDBeFold. Structure Similarity,” n.d.) which resulted in a list of 285 PDB IDs.

The list of PDB IDs with the corresponding RMSD and TM-scores as metrics, and the original PDB extracted list are cross-checked and only the overlapping ID’s are selected and extracted into a new .txt file with the merged ID’s.

The objective of this overlapping cross-check is to enhance the data quality by removing false positives by obtaining these reference structures.

A python script was utilized in order to compare the sequence of the reference dataset with the list of identifiers obtained from the PDB.

2.2 Dataset Cleaning

Starting from the overlapping sequences, redundant sequences were removed with CD-HIT, which is a local sequence alignment based by similarity, the program is user defined performing an all-against-all comparison to obtain two parameters: sequence identity and coverage (length of the alignment) (“CD-HIT. Representative Sequences,” n.d.). For each cluster grouped together, that CD-HIT creates, a representative sequence is taken, typically this could be the one with the longer residue, however this can lose sensitivity so sometimes it may be ideal to choose the one with the one with the highest resolution.

The selected sequences are clustered with CD-HIT which removes redundancy in the set by using the following parameters:

- Default threshold of 90% sequence identity equivalence
- Alignment coverage parameters: aS = 0.8, aL=0.8 (the portion of the sequence that is covered by the alignment)

From this refined set of 26 clustered sequences, the representative structure is manually chosen, which was the first ID of each cluster, and placed into a .fasta file using the first ID of each cluster as reference. This was then cleaned and the first ID of each cluster was placed into a .txt file. A manual selection was done to assess the sequence resolution to ensure those under 3.5A were selected

2.3 Multiple Sequence Structural Alignment

At this point it became possible to start the multiple sequence structural alignment based on the extracted list of PDB identifiers. For this, mTM-Align is used to perform the structural alignment in order to take those selected protein clusters and align them using a multiple sequence alignment (MSA). mTM-Align is an online server software that takes in input a zipped file with the structures in a specific format corresponding to each structure (“TM-align: A protein structure alignment algorithm using TM-score rotation matrix,” n.d.). Structures from the models obtained from the PDB and selected chains were obtained from each using a script, selch.sh. The zipped file of the chains was uploaded to the website, tmAlign and its output of the sequence alignment and the superimposed structural alignment with the scores was stored (Fig 1). Visual analysis of the MSA was done and it was apparent that 1DOD differed particularly from the other sequences, therefore it was removed and the alignment was rerun. The file was saved and cleaned to obtain a more aligned sequence by manually visualizing the areas of the most conserved sequences and constraining the majority of those that do not to obtain a final bpti-kunitz.ali file that will be submitted to HMMER (“HMMER: biosequence analysis using profile hidden Markov models,” n.d.). The results of the MSA along with a manual evaluation of the RMSD are used to build the HMM logo that is created with Skylign (“Skylign. Interactive logos for alignments and profile HMMs,” n.d.).

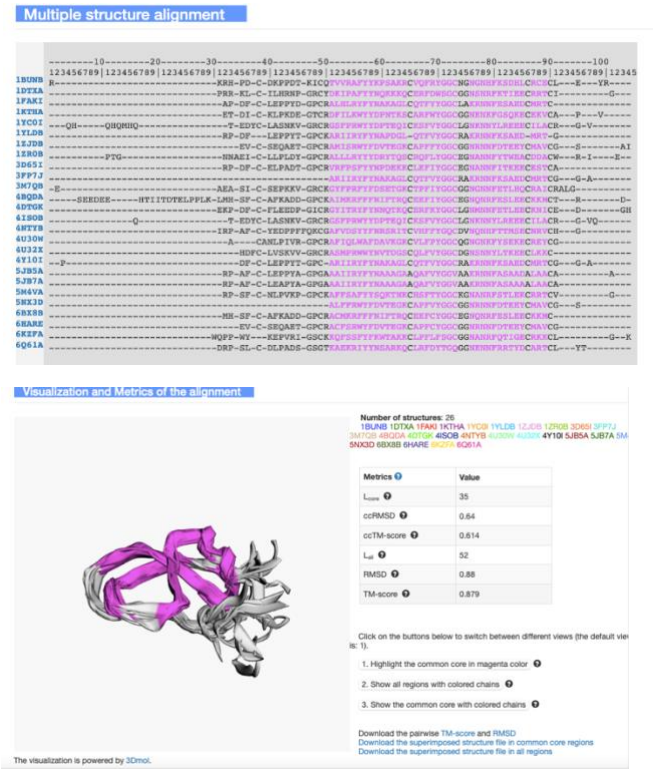

Figure 1. mTM-align output

2.4 Model Generation

Using the hmmbuilt command integrated within HMMER 3.3.2 software, the HMM logo profile was obtained (Eddy, 2011). The results were pictured on the web server Skylign, this logo allowed us to pinpoint the most conserved residues such as Cysteines based on the height of the residues in the logo, this schematic shows the level of conservation which plays a particularly import role as mentioned for this domain (Fig 2).

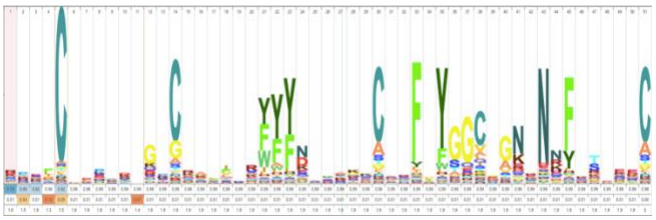

Figure 2. Sequence HMM Profile Logo, Skylign.

2.5 Model Evaluation

To evaluate the quality of the model, a search in the PDB was completed, based on the BPTI Pfam ID PF00014 (Mistry et al., 2021) and specifically searching for those domains by restricting the length of the residues from 40. This search was done to find other proteins potentially with the domain. Those that were reviewed were labelled positives, and then those without review were labelled negative by performing a cross reference including of “AND” and then “NOT” Boolean operators. This search captured a set of 555643 negatives and 359 positives which were then extracted into .fasta files and queried against the bpti-kunitz.hmm file to analyze the values of E and P. Taking the head and tail of each set of positive and negatives, the sets were split randomly and equally into 2 parts, generating four total files of identifiers which were sorted and cleaned according to the necessary columns.

A python script was created to select the list of identifiers from each set and performing a cross validation to begin the optimization of the model.

A hmmsearch using various different options. The same number was used for both the positive and negatives. Option –max was used to remove those with a low seed match, to increase the calculation of the match between the model and the sequence to find only the best match. Parsing of the file was done to understand the positive and negatives, which was done with –noali that removed the alignments to give a cleaner output.

The hmmsearch algorithm returns a list of sequences that were included in the parameters of the range of the score given and only if the hit exists, if not, there will be no output. From this a performance was calculated using python script compare.py and including those sequences that were not present in the hmmsearch output of the training and testing datasets. The Matthews Correlation Coefficient (MCC) and the Accuracy (ACC) is computed for each threshold (Chicco and Jurman, 2020). The best E-value is chosen for both the datasets, based on the higher MCC and ACC. The accuracy, confusion matrix based on precision and are computed. This is repeated for the datasets in a 2-fold cross validation optimizing the threshold for one, and testing the other, and then performing it on the opposing set.

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

2 Results

The HMM proved sufficient ability of extracting the BPTI-Kunitz domain from the multiple structural sequence alignment. A threshold was calculated using the outputs of the MCC, with a sweet point set at 1e-5 for set 1 and 1e-7 for set 2, the accuracy and MCC was high, almost 1, which is what we are looking for. There were 3 false positives and 3 false negatives in total. False positive and false negatives were evaluated in detail and compared to see if it was related to the methods or the annotation. A Clustal Omega search was done on the six misclassified proteins creating an identity matrix (Figure 3), those that were false positives, and those that were false negatives. (“Clustal Omega < Multiple Sequence Alignment < EMBL-EBI,” n.d.).

1: sp|P40500|YII9_YEAST	100.00	21.08	13.33	17.72	17.91	0.00	13.73
2: sp|P78746|CHSD_ASPFU	21.08	100.00	20.65	20.86	17.31	21.31	17.95
3: sp|C5H8E7|TICK1_RHIAP	13.33	20.65	100.00	20.83	15.09	26.00	26.09
4: sp|O62247|BLI5_CAEEL	17.72	20.86	20.83	100.00	56.28	30.43	21.09
5: sp|D3GGZ8|BLI5_HAECO	17.91	17.31	15.09	56.28	100.00	19.57	21.97
6: 3TGI_2|Chain	0.00	21.31	26.00	30.43	19.57	100.00	29.23
7: sp|Q11101|YL15_CAEEL	13.73	17.95	26.09	21.09	21.97	29.23	100.00

Figure 3. Identity Matrix

This identity matrix shows the sequence percentage similarity among the primary sequences taken from the proteins amongst the BPTI-Kunitz domain, chain I.

In regards to the false positives, a BLAST search was run to evaluate the percentage similarity with other proteins that could contain be Kunitz type proteins or at least proteins containing the Kunitz type domains. In fact, one blast search results in an alignment with another protein that was classified as false positive (Figure 4). One reason for the false negative is the size of the alignment and the low E value that was apparent in the hmmsearch output.

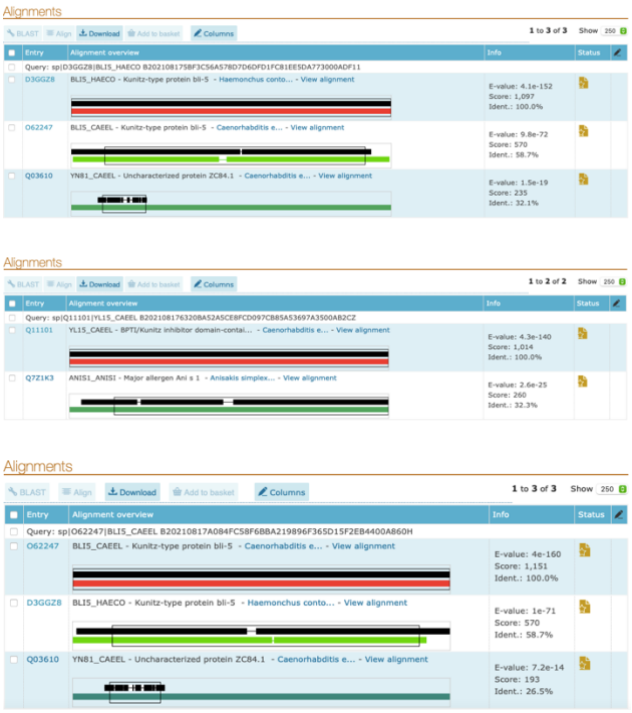

Figure 4. Identity Matrix

In regards to the HMM profile logo, it showed the sequences scanned and those important residues that seemed to be conserved for purposes of protein function and/or stability in the hmmer output were apparent with the 6 conserved cysteines and a tyrosine triplet, this also confirms some similarity to the profile in Pfam.

A confusion matrix was computed as shown in Figure 5a, these computed the actual against the predicted set of positives and negative values obtained. From this binary statistical analysis of classification, the precision and recall were calculated ("sklearn.metrics.f1_score — scikit-learn 0.24.2 documentation," n.d.). This score is the weighted average of the two where the best value is closest or equal one, and the worse closest or equal to zero (Brownlee, 2020).

Precision: 0.99171271

Recall: 0.99171271

In Figure 5b, there is a visual representation of the MCC and ACC scores along with their accuracy. These scores are not to be compared but only visualized graphically the peak of their accuracy in regards to the thresholds.

	Actually positive	Actually Negative
Predicted Positive	359	3
Predicted Negative	3	556002

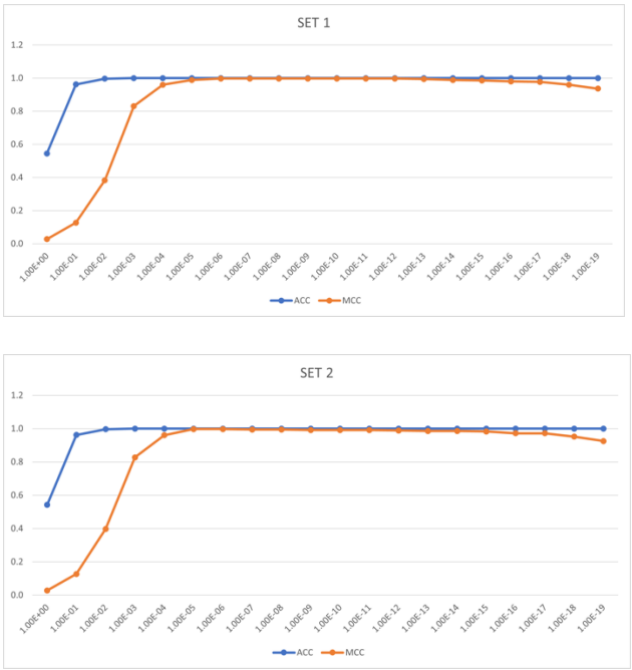

Figure 5. (a) Confusion Matrix with the false positives and negatives; **(b)** Graph of the thresholds.

Overall, the developed method shows reliability for the creation of the Hidden Markov Model to predict the structural for the classification of BPTI/Kunitz domains. For less bias in the output, it could be fruitful to increase the

threshold as sometimes there could be a model included in the validation set that is based on a sequence already present. Achieving a MCC of 1, in these cases are not common, and moreover, they should not be. However, these results prove to be reliable in the fact the false negatives and positives were discovered, along with a solid MCC that confirms the quality of the prediction HMM profile. More importantly the HMM is able to recognize the most conserved regions of the BPTI-Kunitz domains which are the hallmark feature for this domains function and stability.

References

Brownlee, J., 2020. How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Machine Learning Mastery. URL https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ (accessed 8.17.21).

CD-HIT. Representative Sequences, n.d. URL http://weizhong-lab.ucsd.edu/cd-hit/ (accessed 8.16.21).

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6. https://doi.org/10.1186/s12864-019-6413-7

Disulfide Bond | ScienceDirect Topics, n.d. URL https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/disulfide-bond (accessed 8.16.21).

Dittrich, M., Kanchanawarin, C., n.d. Case Study: BPTI.

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLOS Computational Biology 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Clustal Omega < Multiple Sequence Alignment < EMBL-EBI [WWW Document], n.d. URL https://www.ebi.ac.uk/Tools/msa/clustalo/ (accessed 8.17.21).

HMMER: biosequence analysis using profile hidden Markov models, n.d. URL http://hmmer.org/ (accessed 8.16.21).

Karim, S., Adamson, S.W., 2012. Chapter 4 - RNA Interference in Ticks: A Functional Genomics Tool for the Study of Physiology, in: Jockusch, E.L. (Ed.), Advances in Insect Physiology, Small RNAs. Academic Press, pp. 119–154. https://doi.org/10.1016/B978-0-12-387680-5.00004-5

Krokoszynska, I., Dadlez, M., Otlewski, J., 1998. Structure of single-disulfide variants of bovine pancreatic trypsin inhibitor (BPTI) as probed by their binding to bovine β-trypsin11Edited by R. Huber. Journal of Molecular Biology 275, 503–513. https://doi.org/10.1006/jmbi.1997.1460

Kunitz Domain | ScienceDirect Topics, n.d. URL https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/kunitz-domain (accessed 8.16.21).

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Paolo, A., Alessio, B., Martino, B., Andrea, S., Massimo, C., Raimondo, D.C., Enea, M., 2003. The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein. Current Protein & Peptide Science 4, 231–251.

PDBeFold. Structure Similarity, n.d. URL https://www.ebi.ac.uk/msd-srv/ssm/ (accessed 8.16.21).

RCSB Protein Data Bank, n.d. URL https://www.rcsb.org/ (accessed 8.16.21).

Sklearn.metrics.f1_score — scikit-learn 0.24.2 documentation, n.d. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed 8.17.21).

Skylign. Interactive logos for alignments and profile HMMs, n.d. URL https://skylign.org/ (accessed 8.16.21).

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47, D506–D515. https://doi.org/10.1093/nar/gky1049

TM-align: A protein structure alignment algorithm using TM-score rotation matrix, n.d. URL https://zhanglab.ccmb.med.umich.edu/TM-align/ (accessed 8.16.21).