Bioinformatics, 2021, 1–6
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: : 25/10/2021
Scientific article

Laboratory of Bioinformatics I

Building a Profile Hidden Markov Model for the Kunitz-type Protease Inhibitor Domain

Amrou Abas[1,*]

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

Adress: Via San Donato 15, Bologna, Italy

*To whom correspondence should be addressed.

Received on 25.10.2021

Abstract

Motivation: Nowadays, computational approaches regarding functional annotation are greatly sought after due to the accumulation of data regarding protein sequences worldwide. Traditional methods for protein function annotation are mainly based on experiments such as mass spectrometry and microscopy which are considered very time-consuming and resource-demanding due to low throughput and restricted scope of methodology. In this report, a method using a profile Hidden Markov Model (HMM) based on protein structural alignment was achieved to detect a domain that is referred to as the Kunitz domain. In addition, several optimization criteria were used.

Results: A whole dataset was classified to sequences containing the domain and sequences lacking it, achieving accuracy (ACC) and Matthews Correlation Coefficient (MCC) of 0.999 and 0.992 respectively.

Availability: The method is available clicking *here*.

Contact: amrou.abas@studio.unibo.it

Supplementary information: Supplementary data are available clicking *here*.

1 Introduction

The Kunitz domains are known to be the active parts of the protein that prohibit the degredation of other proteins via proteases. Early evidence from basic research and clinical studies shows that the protease inhibitor system is an important factor in inflammatory processes. Inflammation is a key component of the immune system and is associated with atherosclerosis, cardiovascular diseases, diabetes, rheumatoid arthritis, Alzheimer's disease, neurodegenerative disorders, cancer, asthma, and aging[1]. They are relatively small domains of about 50-60 residues long. Some proteins that possess the Kunitz domain include Alzheimer's amyloid prescursor protein (APP), which is a protein that is permanently attached to biological membranes and acts as a cell surface receptor, and it typically consists of 37 to 49 residues, its functions revolve around iron export, synaptic formation, recombination and repair, neuronal plasticity and hormonal regulation. Another Kunitz protein is the polypeptide, whose function is to inhibit Factor Xa, which is an enzyme for coagulation. However, in this report, the focus will be on another protein which is a small one that is called bovine pancreatic trypsin inhibitor (BPTI), also known as aprotinin (Trasylol®). This protein has a role in reducing bleeding during complex surgeries such as heart and liver operations[5].

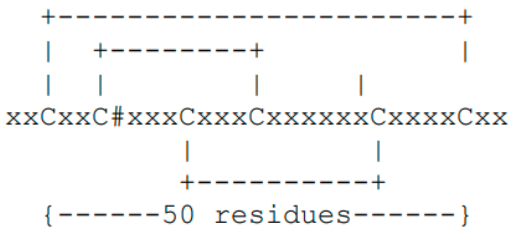


Figure -1-: Kunitz domain: Conserved cysteines involved in a disulfide bond are remarked (C) togheter with active site residue (#) and the position of the pattern (∗).

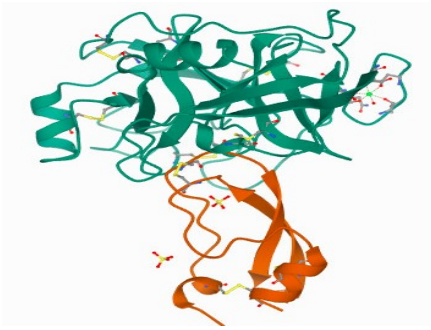


Figure -2-: BPTI structure (PDB entry: 3TGI), in this structure, there is a presence of this protein in a particular functional form, because it is crystallized in the presence of trypsin. The figure includes disulfide bonds (as sticks) which are meant to keep the structure stable[2], it has also a good resolution as a quality of the structure.

BPTI is a monomeric polypeptide composed of 58 amino acids. Its three-dimensional conformation is stable, compact and maintained by three disulfide bridges which control the folding (Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51)[3], and despite the variation in the number of covalent bonds, the spaces between their cysteine residues are conserved[7], conferring their typical disulfide pattern. This inhibitor is found in a limited number of organs, e.g. lungs, spleen, liver and pancreas. It inhibits trypsin, chymotrypsin, kallikrein, and plasmin, with a weak inhibition of human leukocyte elastase, cathepsin G, acrosin and urokinase-type plasminogen activator (uPA)[4]. Usually there are 10 positively-charged lysine (K) and arginine (R) side chains and only 4 negative aspartate (D) and glutamates (E), accounting for the basicity of this protein. Kunitz-type protease inhibitors are found in several organisms including animals, plants, and microbes[6].

A highly exposed active site residue at position 15 is usually lysine (Lys) or arginine (Arg) and is the primary determinant of the specificity of protease inhibition in the way that the long side chain at the aforementioned site binds strongly in the specificity pocket of trypsin, inhibiting its activity, hence, this site is a key factor for the function of the protein.

2 Materials and Methods

2.1 Selection of the initial dataset

As to begin, an identification of the structure that contains the BPTI/kunitz domain is required, and for such matter, the advanced query interface of the Protein Data Bank (PDB)[8] website shall be used, selecting the structures that have good quality, a resolution threshold of 3.5 Å or below is maintained in the query, choosing this threshold because it is the distance to identify two phenomena: hydrogen bond interactions (two electronegative ions that have a hydrogen molecule in the middle can be considered to have a hydrogen bond when the distance is below 3.5 Å), as well as the distance between 2 alpha-carbon atoms, so that if we consider structures that are below 3.5Å, we are almost sure that we are in the range of avoiding possible errors regarding the possibility of presence of hydrogen bonds and also the possibility of mismatch between 2 subsequent alpha atoms.

Another restriction for the query is the size of the protein (since the Kunitz domain has a length between 50 and 60), but we will select a range between 40 and 80, in this way we can avoid the presence of structures that contain multiple domains (if we expect two domains, we should have more than 100 residues). Using it for the aforementioned requirement, the PFAM[9] protein family identifier of the kunitz domain, PF00014, and another filter which is data collection resolution for which a <=3.5 Å threshold will be selected, then using a third filter which is polymer entity sequence length of range 40 to 80 residues (upper incl.). Running the query, 167 structures are obtained. Creating a tabular custom report of the results, using some of the data types that are shown such as PDB ID, entry id (polymer entity identifiers), chain ID, sequence and polymer entity sequence length options.

Generating the report in CSV format (such format was preferred due to the ease of command-line scripts), something to note is that there are still some sequences that are longer than the threshold, because some of them are co-crystallized with chains of other proteins, which obligates us to do a “cleaning procedure” to the file using a bash script.

To verify some low complexity regions at the end of some sequences, we simply type their entry ID into PDB manually, and going to Uniprot[16] from there, we verify that the low complexity region included the kunitz domain because it is still possible that we crystallized a structure that contains the BPTI\Kunitz domain with another structure that has the length of 40 to 80 and we collect them in the dataset which ends in the presence of some negatives here (namely sequences 6HAR, 6BX8 and 4DTG), and they were all manually verified using Uniprot to have included the Kunitz domain. In total, this results in a set of 174 chains, all between 40 and 80 of length containing the Kunitz domain.

As to further check for false positives, we selected from the dataset the 3TGI protein with its I chain and then we went to RUPEE protein structure search[10]. The result had been still verified using PDBe (but still without the ability to actually download the fasta file itself due to another error in the website) when it actually started working, the same list was obtained in both RUPEE or PDBe Fold[11], just to verify because RUPEE is not a very well-known algorithm).

Running a search of that protein against all the proteins that are contained in PDB and according to the level of similarity between this protein and all the proteins in PDB, a subset of proteins with high match will be generated, and these proteins will be selected as possible proteins containing the Kunitz domain, matching these ones with the ones of the first method, in such way that in 2 searches, it is more likely to clean the initial dataset from false positives. Then cleaning and sorting the 2 files to get the entries that are present in both and transporting the result a new file which will have 152 chains out of the first 174. A check for the RMSD and sequence identity shall be conducted on the list of elements that were selected comparing the 3TGI (the one seed protein that was used for the search) against the proteins in the PDB, using a python code, sorting them by increasing RMSD, there is one sequence (1D0DA) that has exceptionally higher RMSD and lower identity scores than the rest (1D0DA stood out at 2.54 Å while the others had ranges closer to one other and varied from 0.11 to 1.74 Å at most). However, accessing PDB and then using the chain to go to the corresponding Uniprot entry, the sequence did indeed include the Kunitz domain, so it will be kept.

As far as one is concerned, there is another issue to deal with now, which is that, many of these structures are redundant (they have structures from proteins that are all the same or very similar). One important point for a good model is the possession of some variability, so we need to look for the redundancy that is present within our dataset, and there are programs that take sequences and clusterize them based on sequence identity. So it is necessary to get the sequences of the common sets of proteins before the clustering because currently we have the PDB ID, but if we need to do the clustering, a comparison of the proteins should be done, creating a file that generates an entry id, the chain (concatenated) and the sequence, generating a fasta file.

A program that is called CD-HIT[12] will be used to clusterize the sequences, which takes a list of sequences and do an all-against-all search, getting back two important results, which are sequence identity and coverage. Sequences based on a level of similarity of such results will be clustered together. We used 90% as a sequence identity cutoff (meaning that, whatever above that percentage will be clustered altogether), and a minimal alignment coverage for the shorter sequence of 0.8. Using this query, a total of 23 clusters were obtained.

Applying the procedure of cleaning the sorted version of the file using a bash script. We then compare the output of the CD-HIT Suite with another clustering algorithm to check for any possible error, a further use of clustering using blastclust shall be used, obtaining 25 clusters. At the end of this step, a representative of the clusters should be chosen, depending on the resolution, a total of 23 seeds were obtained. Now there is the need to go to each one of the PDB files and extract the chain.

2.2 Structural Alignment Generation

To use mTM-align[13], the sequences of the 23 seed representatives must be obtained and extract them from the PDB website, using a basic script that takes as an input a PDB file and a chain identifier and filter out the desired chain, then placing them in a single directory and compressing it. Obtaining the alignment, there are some elements that were not very well-aligned, but the overall alignment is good (figure -4- (a)). However, to make the alignment more compact, the entry 1D0DA.pdb shall be removed from the multiple structural alignment. Also, a 3D depiction of the multiple structural alignment, two matrices of size, one for tm-scores, the other for RMSD values (figure -4- (b)).

Looking at the resultant alignment, RMSD Matrix, the terminal regions are more prone to fluctuations, showing higher RMSD values, and can be more variable with respect to other regions, these parts in the N and C terminal have plenty of gaps and hence shall be removed, but the important aim is to keep the trimming away from the conserved cysteins and at the same time not to include too much gaps.

According to literature, site 15 in BPTI/Kunitz protein is important for the function, checking this from the obtained alignment, that site corresponds to site 21 in the resultant alignment file of mTM-align, and indeed this coincided with what we had mentioned in the introduction that the aforementioned site is often occupied by lysine or arginine.

2.3 HMM Generation using HMMER

After getting the compact alignment. HMMER[14] sequence analysis package will be utilized, and that from which a profile HMM is generated from a Multiple Sequence alignment. Using its *hmmbuild* command, the model shall be generated.

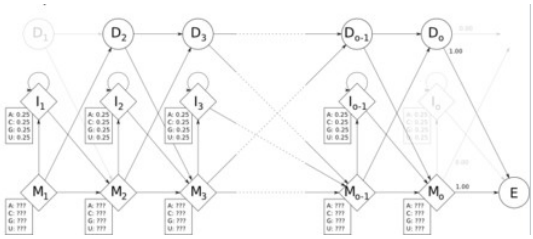


Figure -3-: The profile HMM architecture that HMMER implements.

To visualize the frequency of the residues in the hmm, Skylogo (Skylign)[15] was used, and it was evident that the level of conservation varies across different cysteins, which of course reflects the importance of these particular residues.

2.4 Model Testing

After creating the model, it will be tested against the whole Swissprot[17] part of Uniprot. Downloading the sequences that contain the BPTI/Kunitz domain and whose length is no shorter than 40 residues (356 sequences), and reversing the search query to obtain the negatives which do not identify with this domain (555043 sequences).

To test the model, a cross validation proceduce shall be put in consideration which includes taking a list of identifiers that were possessed in our testing dataset and split it by two, and perform the analysis on one set, identifying the optimal threshold and testing it on the remaining set, which can be done by sorting them randomly and then partitioning each set into two subsets. After preparing the two different sets, there are two options to move on: merge all the positives and negatives in one file and run this file against the model, or running one file at a time. The problem with the second choice is that the sizes of the datasets are different, which affects the e-value (meaning that, in the negative set, the e-value is overboosted because there is a large set), and to overcome this issue, a code using *hmmbuild* shall be used.

After some practical procedures (better explained in the linked Colab web-page) we obtained two sets of all the sequences in Swissprot in which each one has half of them. Now we need to calculate the aforementioned ACC and MCC for the two sets and cross validate the obtained results.

The accuracy (ACC) is the measure of how many predictions are correct on the overall:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}$$

While Matthews Correlation Coefficient (MCC) is a coefficient that assumes 0 for random predictions, 1 for perfect predictions and -1 for completely wrong predictions:

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

The optimization procedure will be better explained in the results section, it is about getting the optimal threshold of one set, applying it on the other, and then doing the same thing vice versa as a form of cross-validation, and then combining the two sets using each others' thresholds.

Obtaining the ID's of the misplaced sequences of two false positives and four false negatives, and searching for their data, the four negatives and one false positive are indeed as what was stated (see results).

3 Results

3.1 Selection of the initial dataset and structural alignment generation

Referring to the selection via PDB advanced search using the three previously mentioned criteria, 167 sequences were obtained and downloaded in csv format, the result is raw data and hence not optimal, so it needed some cleaning, removing double quotes and organizing the file and then removing the co-crystallized sequences as well.

Another test that was made is running a pairwise 3D alignment of 3TGI chain I against the PDB using RUPEE (using default settings) and matching the common ones between that and the method mentioned above resulted a list that is clear from possible errors that could propagate, getting 152 PDB entry IDs concatenated to a specific chain identifier, as it seems that some sequences of the initial 174 do not have similar structures, the 152 will be co-located in a file.

After clusterizing them to get variability, we still need the fasta format to get it accepted by CD-HIT cluster generator. Getting the result of 23 clusters containing the 152 sequences (all default but the identity cutoff of 90% and minimal alignment coverage for shorter sequences of 0.8), and to make sure everything is correct, another clustering is done using blastclust with a bash code with the same parameters as in CD-Hit, yielding 25 clusters.

Then, a seed representative of each cluster was chosen on the basis of best resolution, resulting in 23 seeds in file that contains their identifiers. now we have generated a list of PDB files containing all the specific chains of the proteins we are trying to align, and compressing their file directory to be inserted onto mTM-align interface and doing multiple structural alignment. Using mTM-align with default settings, the result of the multiple structural alignment is as follows (figure -4- (a)).

3.2 HMM Generation using HMMER

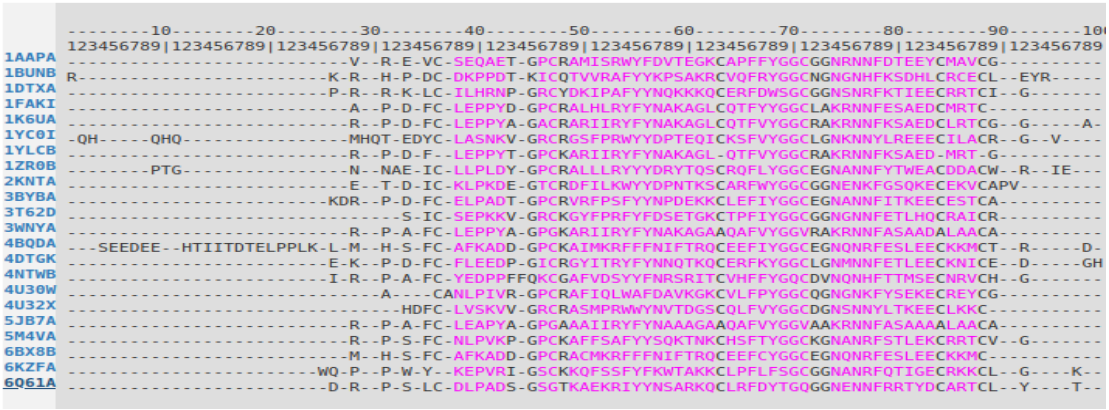
Cleaning and organizing the file, cutting the gapped edges of the previously shown alignment from position 1 to 28 and from 61 till the end and then making it in a fasta format to form the model generating the HMM with the following HMM logo (figure -5- (c)).

3.3 Model Testing: Obtaining two sets and Optimization procedure

Testing the efficiency and the reliability of this model through testing it against all Swissprot entries, dividing them in two sets, each one including half the total number of negatives and positives, and then testing the performance of our the result for two sets (that includes all positives and negatives) using a python script.

We have concluded that the best performance of set one was recorded at 1e-05. Now, applying this threshold on set two, the result will be at threshold 1e-05 with ACC and MCC of 0.999 and 0.997 respectively. Then taking the optimal threshold at 1e-11 of set two and applying it to set one and getting ACC and MCC of 0.999 and 0.992 respectively, as a form of cross-validation. Creating a set that gathers both these 2 sets using a bash script applying the thresholds of set one to two and vice versa, obtaining a total set that contains all the results and their classification according to the model (that is, whether they are false

a)



b)

Metrics	Value
L_{core}	42
ccRMSD	0.73
ccTM-score	0.698
L_{ali}	54
RMSD	0.92
TM-score	0.883

C)

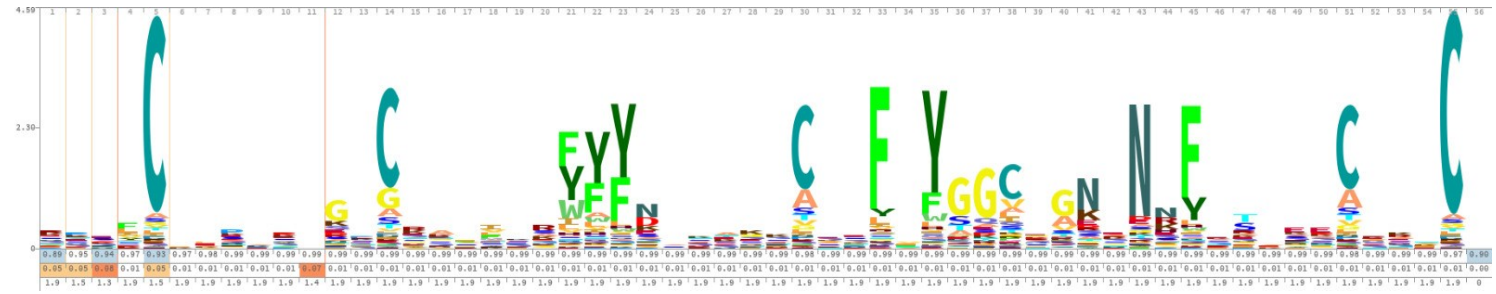


Figure -4-: a- Structural alignment using the 22 chosen seed sequences (after removing 1D0D with its chain A because it was not very well aligned, the purple part is very well aligned with respect to the other that show fluctuations), b- The RMSD and TM-Score matrices are available (rmsd-score.matrix and tm-score.matrix in supplementary materials). c- Using Skylign Logo to create a visualization of the sites represented by each position, the important cysteins are evident to be very well conserved but with varying degrees, and in bottom picture, the contents of bpti-kunitz.ali where it is seen that the position important to the function is well conserved (position 21 here, position 15 in literature, mostly either lysine or arginine).

positives, true negatives, etc), also we can get the performance of the full dataset using a python script with the threshold of 0.5 (choosing it because we need to a division between zero and one because what is above 0.5 goes with the positives and what is below it goes with the negatives) which will result in 0.999 and 0.992 as ACC and MCC respectively.

3.4 Model Testing: Confusion Matrix

Using a bash script to get the number of false positives and negatives, the result included two false positives and four false negatives while the rest were classified correctly (table -1-). The model showed good performance in general and classified the sequences in a correct way except for some few sequences that are going to be investigated.

		Actual Class	
		Condition Positive	Condition Negative
Predicted Class	Test Outcome Positive	True Positive 352	False Positive 2
	Test Outcome Negative	False Negative 4	True Negative 555041

Table-1-: Confusion matrix of the full dataset.

3.5 Model Testing: Discussion on misclassified proteins

The whole approach resulted in four false negatives that were:

i- D3GGZ8 (Kunitz-type protein bli-5): Running the sequence in a fasta format against our model (file D3GGZ8.fasta), no hits (domains) were obtained satisfying reporting thresholds, searching in literature, turns out it was due to the absence of key residues from its equivalent bovine pancreatic trypsin inhibitor motif (Fx(3)GCx(6)FYx(5)C)[18][19]; but it is still uncertain as bli-5 lacks all the catalytic features of serine proteases.

ii- P86963 (BPTI/Kunitz domain-containing protein 2): Running it against our model, it gave a hit for three separate domains, the reason behind it being false negative is because two of the three domains have a high e-value (0.013 and 0.8) but indeed checking it had protease inhibitor activity and the third domain has a good e-value (1.1e-09) with all most occurring cysteins present except one in the alignment and it was classified as BPTI/Kunitz via a rule (PROSITE-ProRule:PRU00031) in Uniprot, and it is constrained with three disulfite bonds as well.

iii- Q11101 (BPTI/Kunitz inhibitor domain-containing protein C02F12.5): Running it against our model, it gave one domain with 4.4e-09, but aligned from 16 to 55 in our model so the two first important cysteins of the most occurring in our model (see Skylogo) were not in the aligned domain. Looking through its Uniprot entry, it indeed included 3 disulfide bonds and was identified as Kunitz via PROSITE-ProRule:PRU00031.

iv- O62247 (Kunitz-type protein bli-5): Same case as D3GGZ8.

And two false positives that were:

v- P56409 (Ornithodorin): A highly selective thrombin inhibitor, running the sequence against the model, three domains were obtained, two of them have high e-value but the third has a low e-value with all the important cysteins conserved, it was wrongly included in the negative dataset, because on Uniprot it is recognized as Kunitz but lacked the PFAM identifier (PF00014) which was used to build this dataset.

vi- P40500 (Uncharacterized membrane protein YIL089W): Running it against the model, no hits were detected for any domain, no presence for any BPTI/Kunitz domain on its Uniprot page or even in the IDs that are similar to it, on PFAM and InterPro it is classified as protein of unknown function, hence its inclusion should be due to a technical mistake related to the model itself.

4 Conclusion

A good model for detecting the presence of potential uncharaterized BPTI/Kunitz domains was generated using seed sequences of 22 sequences that were the representatives of each cluster then aligning them and producing the model using HMMER, showing very high ACC, MCC and very low number of mismatched sequences. This approach can also be implemented for other domains that may play vital roles in the biomedical sector to identify and hence annotate novel sequences as the use of HMM of identifying protein families becomes more and more of a necessity in the scientific community.

Acknowledgements

Thanks to all professors who took their time to answer questions and to my colleagues who always engaged in stimulating discussions which have often led to new perspectives and conclusions.

References

1- Ferencík M, Stvrtinová V, Hulín I, Novák M. Inflammation–a lifelong companion: attempt at a non-analytical holistic view. Folia Microbiol (Praha). 2007;52(2):159–73.

2- Vincent JP, Lazdunski M. Trypsin-pancreatic trypsin inhibitor association: dynamics of the interaction and role of disulfide bridges. Biochemistry. 1972;11(16):2967–77.

3- Kassell B, Laskowski M. The basic trypsin inhibitor of bovine pancreas. V. The disulfide linkages. Biochemical and Biophysical Research Communications. 1965;20(4):463–8.

4- Mahdy AM, Webster NR. Perioperative systemic haemostatic agents. British Journal of Anaesthesia. 2004;93(6):842–58.

5- Flight SM, Johnson LA, Du QS, Warner RL et al. (2009)Textilinin-1, an alternative anti-bleeding agent to aprotinin: importance of plasmin inhibition in controlling blood loss. British Journal of Haematology. 2009;145:207–211.

6- Rawlings ND, Tolle DP. Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem Journal. 2004;378:705–716.

7- Pritchard L, Dufton MJ. Evolutionary Trace Analysis of the Kunitz/BPTI Familyof Proteins: Functional Divergence May Have BeenBased on Conformational Adjustment. 1999;285:1589-1607.

8- Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB):
ensuring a single, uniform archive of PDB data. Nucleic Acids Research. 2007;35:301–303.

9- Bateman *et al.* The Pfam protein families database. Nucleic Acids Research. 2004;32:138-141.

10- Ayoub R, Lee Y. RUPEE: A fast and accurate purely geometric protein structure search. 2018.

11- Gutmanas *et al*. PDBe: Protein Data Bank in Europe. 2013;42:285-291.

12- Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing
biological sequences. 2010;26:680-682.

13- Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. 2018;34(10):1719-1725.

14- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. 2011;39:29-37.

15- Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. 2014, 15:7.

16- The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. 2020;49:480-489.

17- Boeckmann *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003;31:365-370.

18- Stepek G, McCormack G, Page AP. The kunitz domain protein BLI-5 plays a functionally conserved role in cuticle formation in a diverse range of nematodes. 2010;169:1-11.

19- Skuce *et al*. Cloning and characterisation of thrombospondin, a novel multidomain glycoprotein found in association with a host protective gut extract from Haemonchus contortus. 2001;117:241-244.