Report
Submission due to May 2022
Alma Mater Studiorum –University of Bologna

Laboratory of Bioinformatics

Development and implementation of a profile Hidden Markov Model to detect Kunitz-BPTI domain

Vuong Viola Meixian[1]

[1]Department of Pharmacy and Biotechnology, LM in Bioinformatics, University of Bologna, Italy

Last revision May 15th, 2022

Abstract

Motivation: The BTPI-Kunitz family is one of the most extensively studied protein family, whose members are mainly involved in the cleavage of serine proteases. Due to their abundance, stability and versatility, these proteins have been exploited for the investigation of a large variety of protein systems. Therefore, given their importance, a profile Hidden Markov model, modelled from the Kunitz domain, is presented in this study.
Results: The results confirm how Hidden Markov models are powerful tools able to detect and correctly differentiate Kunitz-like proteins from non-Kunitz ones, with only a small margin of error.
Contact: violameixian.vuong@studio.unibo.it
Supplementary information: supplementary materials are available.

1 Introduction

The Kunitz-type protein family (Pfam accession ID: PF00014) is a large family of serine protease inhibitors consisting of over 200 sequences. Members of this family – known also as I2 – has been discovered in multiple species across different domains, from eukaryotes, bacteria, to viruses (Rawlings *et al.*, 2004).

All Kunitz-type proteins, with some exceptions, share the same inhibitory mechanism, the ‘standard’ mechanism, which was first described by Laskowski and Qasim (Laskowski and Qasim, 2000). According to their study, cleavage of serine proteases is triggered following a high-affinity binding between the active site of the enzymes and the Kunitz-domain of the inhibitor. This will lead to a series of conformational changes, allowing the resulting complex to be recognized and cleared from the circulation. The study also confirms how this mechanism can only be observed in this protein family (Rawlings *et al.*, 2004).

By definition, the protein function is directly correlated to its three-dimensional structure. Nevertheless, all Kunitz-type proteins possess the same scaffold. A better insight regarding this matter can be addressed describing the representative of the family, the well-known bovine pancreatic trypsin inhibitor BPTI.

BPTI is a 58 amino-acid long globular protein, characterized by a single Kunitz-inhibitor domain, a low relative molecular mass, and a basic isoelectric point (Kunitz and Northrop, 1934). All Kunitz-type proteins share these same characteristics with one or several domain(s) and a broad spectrum of activity towards serine proteases (Barrett and Salvesen, 1986). The three-dimensional structure of BPTI was first crystallized by Huber *et al*. and consists in α-helixes and antiparallel β-sheets, connected through loops, and stabilized by three disulphide bridges (Huber *et al*, 1970). Their importance is also linked to the phylogenetic origin of the domain. Since Kunitz-type proteins have evolved so much in respect to the ancestral Kunitz-type domain, nowadays members of the family can be discriminate by assessing the presence of these disulphide bridges (Kunitz and Northrop, 1934).

The Kunitz domain is a relatively small domain, typically 50-60 residues long. Six cysteines, identically spaced, can be found in all BPTI-like domains, along a peptide chain of varying length, as generally indicated below (with the number of amino acids between consecutive cysteines in parentheses):

$$H_2N - \ldots - Cys - (8) - Cys - (15) - Cys - (7) - Cys - (12) - \\ - Cys - (3) - Cys - \ldots - CO_2H$$

The presence of these cysteines is fundamental to the formation of the disulphide bridges. Due to their stabilization role in the tertiary structure of BPTI-like proteins, it is not strange that they are highly conserved.

Since the discovery of BPTI, the importance of BPTI-like proteins has since been demonstrated. The stability and abundance of BPTI and all BPTI-related domains have made them a convenient model protein for experimental and theoretical studies. They can in fact be covalently inserted into numerous large protein systems, as diverse as human blood coagulation and Alzheimer’s disease (Carrel, 1988; Ponte *et al.*, 1988). Therefore, knowledge of both their inhibitory mechanism and structure can be exploited for possible therapeutic applications.

The aim of this study is to build a Hidden Markov Model (HMM) that is able to set apart Kunitz-type proteins from non-Kunitz-type ones. The model will be build using as seed known Kunitz-protein structures, followed by a statistical assessment of its performance. A critical discussion of the results will be also provided to establish the ability of the model itself to correctly classify SwissProt Kunitz and non-Kunitz protein sequences.

2 Methods

2.1 Data collection

Kunitz-type protein structures were retrieved from two different sources – PDB (Burley *et al.*, 2021) and PDBeFold (Krissinel and Henrick, 2004). The first set of Kunitz structures was obtained exploiting the advanced search of PDB. The selection was executed following these Structure Attribute criteria *i)* Identifier - Pfam Protein Family equals PF00014, *ii)* Data Collection Resolution <= 2, *iii)* Polymer Entity Mutation Count = 0 (supplementary material S1).

The second search was performed on PDBeFold. The known Kunitz/BPTI structure 3TGI (chain I) was used as query to recover all its matches above 70% (specified in the parameter lowest acceptable match (%)) found in the whole PDB archive (supplementary material S2).

2.2 Data cleaning and manipulation

The raw dataset files obtained were manipulated to detect all the overlapping PDB IDs and their relative chain(s). The aim is to obtain a single, clean list of Kunitz structures for the following analyses. In particular, the custom table of entries fetched from PDB was handled with a Python script to fill out all those blank spaces which were unfortunately present (supplementary material S3).

Among all structures retrieved following the comparison, however, there were highly similar sequences, if not identical. Therefore, redundancy was removed launching the web server CD-HIT (Huang *et al*, 2010). As the software takes as input FASTA sequences, a priori step to retrieve all PDB IDs FASTAs was required. CD-HIT was then run using as a sequence identity cut-off 0.99; all remaining parameters were left in the default setting. From each resulting cluster, the representative ID (the one which is the longest, as defined by CD-HIT) was collected for the following multiple structural alignment step.

2.3 Multiple structural alignment (MSA) and construction of a profile Hidden Markov Model (HMM)

The multiple structural alignment was performed with PDBeFold v2.59 (Krissinel and Henrick, 2005) using as query the PDB IDs of the clusters representative previously obtained from CD-HIT. From this multiple structural alignment, the multiple sequence alignment (MSA) was retrieved (supplementary material S4) and graphically visualized with MView 1.63 (Madeira *et al*, 2022). The MSA was then used as query to build a profile HMM using the option hmmbuild of HMMER 3.3.2. (Eddy, 2009) (supplementary material S5). The model logo was then generated using the software Skylign (Wheeler *et al*, 2014).

2.4 Model training and testing

To train and validate the model, a set of positive (containing the Kunitz-domain) (supplementary material S6) and a set of negative (not containing the Kunitz-domain) (supplementary material S7) sequences were both retrieved from SwissProt (UniProt Consortium, 2021). The queries below, respectively for positives and negatives, were used to perform the advanced search in Uniprot:

database:(type:pfam pf00014) length:[40 TO 1000] NOT database:(type:pdb) AND reviewed:yes

NOT database:(type:pfam pf00014) length:[40 TO 1000] reviewed:yes

Following, a 2-fold cross validation of the model was performed. Both sets – positives and negatives – were therefore split in half. To avoid introducing bias during the splitting process, each set division was performed after shuffling randomly the UniProt IDs. For each of these subsets UniProt ID, a Python script was run (supplementary material S8) to retrieve the corresponding FASTA sequence. Each resulting file, along the Kunitz model previously build, was used as input to execute HMMER hmmsearch. The following parameters, *i)* --$noali$, to not show the alignment, *ii)* -- max, to turn all heuristic filters off (the algorithm is more powerful, but slower), *iii)* -Z 1, to normalize the E-value calculation, *iv)* --$domZ$ 1, to normalize the domain E-value calculation, *v)* -$tblout$, to arrange in tabular form the results, were selected to secure more refined and subset size-independent results. As the hmmsearch algorithm failed to retrieve all negatives, those which are missing were tracked down by searching them in the negative set. From each tabular output generated from hmmsearch, the UniProt ID, its corresponding E-value domain, and its class – either 1 for positives or 0 for negatives – were extracted. We then combined the first positive and negative subsets into a unique 'first' set (here, set_1); the same procedure was applied for the second positive and negative subsets (set_2).

2.5 Model assessment and optimization

The statistical evaluation of our model was performed running a Python script (supplementary material S9). The program returned as output the threshold set as input, the accuracy ACC (1), the Matthews Correlation Coefficient MCC (2), and the confusion matrix containing the number of true positives, false positives, false negative, true negative, in this order.

$$ACC = \sqrt{\frac{TN + TP}{TN + TP + FN + FP}} \quad (1)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

with TN: # of true negative; TP: # of true positive; FN: # of false negative; FP: # of false positive.

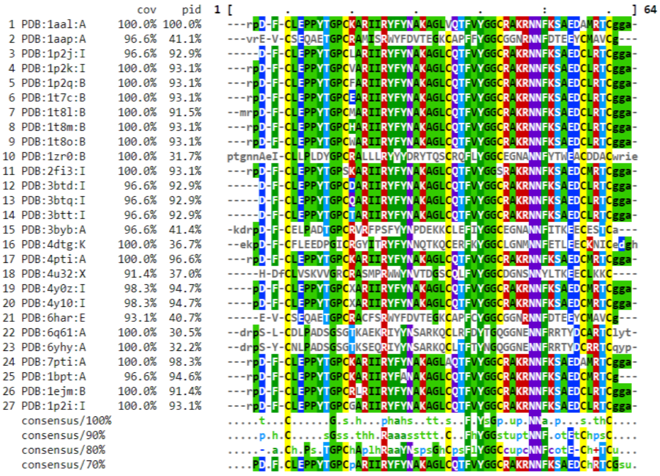
To find the optimal threshold for each set of sequences, a range of possible threshold values was selected. The range covered the interval between 1e-3 and 1e-10, scaling by an order or magnitude. The E-value threshold which maximize best the MCC value was selected; if more than one threshold shared the same MCC value, their mean was selected as the optimal one.

Two separate runs were carried out – one for set_1, the other for set_2 – and another run was executed for the entire dataset using as threshold the mean between the two runs' thresholds. From this execution, the confusion matrix, ACC, MCC were evaluated, as well as the FP and FN, which were further investigated by retrieving any information regarding their classification in UniProt (UniProt Consortium, 2021).

3 Results

160 and 648 structures were retrieved respectively from PDB and PDBeFold. Upon comparison, only 112 structures were found belonging to both datasets. CD-HIT grouped these 112 sequences into 27 clusters, from which the representative structures were retrieved.

The structural and sequential alignment of the 27 representatives (Fig. 1) returned an overall RMSD and Q-score of 0.5786 and 0.7961, respectively. For each representative the corresponding RMSD and Q-score value, as well as the number of residues and secondary structure elements (SSEs) are reported (Table 1).



```
                            cov      pid   1 [        .         .         .         .         :         .    ] 64
 1 PDB:1aal:A           100.0%  100.0%    ---rpD-F-CLEPPYTGPCKARIIRYFYNAKAGLVQTFVYGGCRAKRNNFKSAEDAMRTCgga-
 2 PDB:1aap:A            96.6%   41.1%    ---vrE-V-CSEQAETGPCRAMISRWYFDVTEGKCAPFFYGGCGGNRNNFDTEEYCMAVCg---
 3 PDB:1p2j:I            96.6%   92.9%    -----D-F-CLEPPYTGPCLARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 4 PDB:1p2k:I           100.0%   93.1%    ---rpD-F-CLEPPYTGPCVARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 5 PDB:1p2q:B           100.0%   93.1%    ---rpD-F-CLEPPYTGPCFARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 6 PDB:1t7c:B           100.0%   93.1%    ---rpD-F-CLEPPYTGPCEARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 7 PDB:1t8l:B           100.0%   91.5%    --mrpD-F-CLEPPYTGPCMARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 8 PDB:1t8m:B           100.0%   93.1%    ---rpD-F-CLEPPYTGPCHARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
 9 PDB:1t8o:B           100.0%   93.1%    ---rpD-F-CLEPPYTGPCWARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
10 PDB:1zr0:B           100.0%   31.7%    ptgnnAeI-CLLPLDYGPCRALLLRYYYDRYTQSCRQFLYGGCEGNANNFYTWEACDDACwrie
11 PDB:2fi3:I           100.0%   93.1%    ---rpD-F-CLEPPYTGPSKARIIRYFYNAKAGLCQTFVYGGSRAKRNNFKSAEDCMRTCgga-
12 PDB:3btd:I            96.6%   92.9%    -----D-F-CLEPPYTGPCDARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
13 PDB:3btq:I            96.6%   92.9%    -----D-F-CLEPPYTGPCQARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
14 PDB:3btt:I            96.6%   92.9%    -----D-F-CLEPPYTGPCTARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
15 PDB:3byb:A            96.6%   41.4%    -kdrpD-F-CELPADTGPCRVRFPSFYYNPDEKKCLEFIYGGCEGNANNFITKEECESTCa---
16 PDB:4dtg:K           100.0%   36.7%    --ekpD-F-CFLEEDPGICRGYITRYFYNNQTKQCERFKYGGCLGNMNNFETLEECKNICedgh
17 PDB:4pti:A           100.0%   96.6%    ---rpD-F-CLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCgga-
18 PDB:4u32:X            91.4%   37.0%    -----H-DfCLVSKVVGRCRASMPRWWYNVTDGSCQLFVYGGCDGNSNNYLTKEECLKKC----
19 PDB:4y0z:I            98.3%   94.7%    ----pD-F-CLEPPYTGPCXARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCgga-
20 PDB:4y10:I            98.3%   94.7%    ----pD-F-CLEPPYTGPCXARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCgga-
21 PDB:6har:E            93.1%   40.7%    -----E-V-CSEQAETGPCRACFSRWYFDVTEGKCAPFCYGGCGGNRNNFDTEEYCMAVCg---
22 PDB:6q61:A           100.0%   30.5%    --drpS-L-CDLPADSGSGTKAEKRIYYNSARKQCLRFDYTGQGGNENNFRRTYDCARTClyt-
23 PDB:6yhy:A           100.0%   32.2%    --drpS-Y-CNLPADSGSGTKSEQRIYYNSARKQCLTFTYNGQGGNENNFRRTYDCRRTCqyp-
24 PDB:7pti:A           100.0%   98.3%    ---rpD-F-CLEPPYTGPCKARIIRYFYNAKAGLAQTFVYGGCRAKRNNFKSAEDAMRTCgga-
25 PDB:1bpt:A            96.6%   94.6%    ---rpD-F-CLEPPYTGPCKARIIRYFANAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCg---
26 PDB:1ejm:B           100.0%   91.4%    ---rpD-F-CLEPPYTGPCRLRIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
27 PDB:1p2i:I           100.0%   93.1%    ---rpD-F-CLEPPYTGPCGARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCLRTCgga-
   consensus/100%                         .....t...C......G.s.h...phahs..tt.s..F.YsGp.up.NNa.p...s.thC....
   consensus/90%                          .....p.h.C.....sGss.thh.Raaassttt.C..FhYGGstuptNNF.otEtChpsC....
   consensus/80%                          .....-.a.Ch.Ps.TGPChAplhRaaYNspsGhCpsFlYGGCcupcNNFcotE-Ch+TCu...
   consensus/70%                          ....PD.F.CLEPPYTGPCpARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDChRTCGsu.
```

Fig. 1. Multiple sequence alignment of the 27 representative sequences. In each position of the alignment, the most conservative residue is highlighted. Reported here are also the coverage, percentage identity and consensus sequence at different percentage identity.

From each position of the alignment, the most conservative residue has been highlighted. We can observe how all six cysteines are highly conserved in most of the sequences. Each representative has been reported along its coverage (cov) and percentage identity (pid) with respect to the reference 1aal:A. The overall coverage is quite good, always above 90%, while 19 out of 27 sequences match the reference at least 90%, suggesting their high similarity.

Table 1. Representatives' multiple structural alignment scores. RSMD indicates how good is the superposition of our structures; Q-score indicates how many residues in equivalent SSE are superimpose well in three-dimensional space.

			Consensus scores	
Structure	# of residues	# SSE	RMSD	Q-score
1aal:A	58	4	0.2739	0.9062
1aap:A	56	4	0.4894	0.9219
1bpt:A	56	4	0.3075	0.9366
1ejm:B	58	4	0.2523	0.9074
1p2i:I	58	4	0.2361	0.9082
1p2j:I	56	4	0.2620	0.9393
1p2k:I	58	4	0.2453	0.9077
1p2q:B	58	4	0.2515	0.9074
1t7c:B	58	4	0.2395	0.9080
1t8l:B	59	4	0.2598	0.8916
1t8m:B	58	4	0.2417	0.9079
1t8o:B	58	4	0.2462	0.9077
1zr0:B	63	4	0.6383	0.8048
2fi3:I	58	4	0.2415	0.9079
3btd:I	56	4	0.2337	0.9407
3btq:I	56	4	0.2289	0.9409
3btt:I	56	4	0.2257	0.9411
3byb:A	58	4	0.6208	0.8763
4dtg:K	60	4	0.5679	0.8528
4pti:A	58	4	0.3096	0.9042
4u32:X	54	3	0.4985	0.9551
4y0z:I	57	4	0.2230	0.9247
4y10:I	57	4	0.2257	0.9246
6har:E	54	4	0.6846	0.9329
6q61:A	59	4	0.6867	0.8536
6yhy:A	59	4	0.7306	0.8480
7pti:A	58	4	0.3695	0.9001

The multiple sequence alignment was then fed to the HMMER hmmbuild algorithm, from which our profile HMM was build. The model's logo (Fig.2), provided by Skylign, shows the graphical representation of our 58-residues long profile. In each position, a stack of letters is drawn: the stack's height corresponds to the conservation at that position, while each letter's height (represented by a unique colour) depends on the frequency of that letter at that position. We can observe that cysteines rule positions 5, 14, 30, 38, 51, and 55, with C5 and C55 being the maximum observed height (represented on the y-axis).

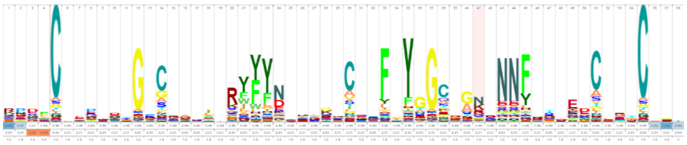


Fig. 2. Profile Hidden Markov model of the Kunitz domain logo. The maximum value on the y-axis is the Maximum Observed Height (in bits); displayed on the x-axis are occupancy, insert probability, expected insert length, in order.

For the 2-fold cross-validation of our profile HMM, 323 positives and 538 623 negatives were retrieved from UniProt. The hmmsearch launched on the negative set failed to recover 281 714 sequences, which were acquired after a comparison search.

The Python pipeline set to search the optimal threshold returned a value of 1e-5 and 1e-7 for set_1 and set_2 respectively. Each of these values, applied on the opposite set, returned MCC values of 0.98 and 0.99. The mean between these two thresholds – corresponding to our overall threshold for the entire dataset – is 1e-6. From it, the confusion matrix (Table 2) was computed. Both ACC and MCC values scored 0.99.

Table 2. Confusion matrix.

	Actual positive (1)	Actual negative (0)
Predicted positive (1)	TP (1, 1) = 321	FP (1, 0) = 5
Predicted negative (0)	FN (0, 1) = 2	TN (0, 0) = 538 618

Five FP and two FN were discovered upon investigation (Table 3).

Table 3. E-value of both FN and FP.

	UniProt ID	E-value
FN	D3GGZ8	1e-05
	O62247	1.2e-06
FP	P0DV03	1e-27
	P0DV04	6.2e-27
	P0DV05	1.3e-28
	P0DV06	9.6e-27
	P56409	5.1e-7

4 Discussion

Since BPTI discovery, several X-ray crystallography and NMR studies have been conducted on Kunitz-like proteins (Deisenhofer and Steigemann, 1975; Ponte et al, 1988; Berndtm, 1993). Through these studies, the highly conservation of the six cysteines distributed along the domain have been since demonstrated. Nevertheless, these residues are involved in the formation of three disulphide bridges (Cys5 – Cys55, Cys14 – Cys38, Cys30 – Cys51) which are involved in the overall stabilization of the tertiary conformation (Huber *et al*, 1970). Our MSA and profile HMM logo confirm this conservation pattern.

These conformational studies also demonstrate how Kunitz-type proteins share the same scaffold, consisting in one or two helical regions and two antiparallel β-sheets. The MSA conducted in our experiments shows that all structures are characterized by four SSE, with the exception of 4u32:X, which possess just a single α-helix (Pendlebury *et al.*, 2014). In particular, the second sheet extends up to residue 35, which is followed by the tripeptide segment Gly36 – Gly37 – Cys38. Sequence alignment conducted on all SwissProt Kunitz-type inhibitors shows that the residue preceding Cys38 is always a strictly conserved glycine (Antuch *et al*, 1994). This result is also confirmed in our study. Looking at our model logo, both Gly36 and Gly37 are the highest letter, meaning the probability of encounter a Gly in these positions is high (probability frequency of 0.432 and 0.825, respectively). The conservation of Gly37, especially, is fundamental, as the local backbone conformation in BPTI, and therefore BPTI-like 3D structures, can only be satisfied by this residue (Antuch *et al*, 1994).

We therefore conclude that the profile HMM we built generalize well the peculiar features of the Kunitz domain. Nevertheless, our profile model seems to predict with a certain degree of accuracy Kunitz and non-Kunitz proteins. Out of 538 946 sequences, only 7 were misclassified.

Evidence on UniProt demonstrate the membership to the Kunitz family of false negatives O62247 and D3GGZ8. Both proteins aren’t involved in any serine protease inhibition mechanism, but their participation in the cuticle biosynthesis has been proven since (Stepek *et al*, 2010; Page *et al*, 2006). A multiple sequence alignment performed with Align (Sievers *et al.*, 2014) of these two proteins against the pancreatic trypsin inhibitor P00974 was conducted. An identity of 7% confirms how both proteins are not similar enough to the representative proteins from which our profile HMM was built, from which the inability of our model to detect them as Kunitz proteins derived.

The low annotation score and the lack of a Pfam annotation may represent the motive of the misclassification of false positive P56409. Nevertheless, it seems that it has a serine protease inhibition function, and two BPTI domains. Still, no strong evidence supports this statement. The other false positives – P0DV05, P0DV03, P0DV04 and P0DV06 – are, on the other hand, classified as non-Kunitz proteins. However, proofs found on UniProt regard them as Kunitz proteins, belonging specifically to the venom Kunitz-type protein family. The sequences share a 75% identity with P31713, a known Kunitz-family member from which they were inferred from; it’s also important to specify that all the cysteines involved in the disulphide bridge formation are conserved. The positive prediction formulated by our model seems to be in accordance with these evidence, but the lack of a Pfam annotation will still regard them as negatives. As these UniProt entries are still fairly new (uploaded date: February 2022), we expect changes regarding their annotation in the near future.

References

Antuch, W., Güntert, P., Billeter, M., Hawthorne, T., Grossenbacher, H. and Wüthrich, K. NMR solution structure of the recombinant tick anticoagulant protein (rTAP), a factor Xa inhibitor from the tick Ornithodoros moubata, FEBS Letters, 1994;352.

Barrett, A. J. and Salvesen, G. Proteinase inhibitors / editors, A.J. Barrett and G. Salvesen. Elsevier Amsterdam; New York, 1986.

Berndt, K.D, Gunter, P. and Wuthrich, K. J. Mol Biol. 1993;234,735-750.

Burley SK, Bhikadiya C, Bi C, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 2021;49(D1):D437-D451.

Carrell, R. Enter a protease inhibitor. Nature 1988;331,478–479.

Deisenhofer, J. & Steigemann, W. Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1.5Å resolution. Acta Cryst. 1975;B31,238-250.

Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23(1):205-211.

Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680-682.

Huber R, Kukla D, Rühlmann A, Epp O, Formanek H. The basic trypsin inhibitor of bovine pancreas. I. Structure analysis and conformation of the polypeptide chain. Naturwissenschaften. 1970;57(8):389-392.

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004;60(Pt 12 Pt 1):2256-2268.

Krissinel, E., Henrick, K. (2005). Multiple Alignment of Protein Structures in Three Dimensions. In: R. Berthold, M., Glen, R.C., Diederichs, K., Kohlbacher, O., Fischer, I. (eds) Computational Life Sciences. CompLife 2005. Lecture Notes in Computer Science(), vol 3695. Springer, Berlin, Heidelberg.

Kunitz M, Northrop JH. THE ISOLATION OF CRYSTALLINE TRYPSINOGEN AND ITS CONVERSION INTO CRYSTALLINE TRYPSIN. Science. 1934;80(2083):505-506.

Laskowski M, Qasim MA. What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes?. Biochim Biophys Acta. 2000;1477(1-2):324-337.

Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022;gkac240.

Page AP, McCormack G, Birnie AJ. Biosynthesis and enzymology of the Caenorhabditis elegans cuticle: identification and characterization of a novel serine protease inhibitor. Int J Parasitol. 2006;36(6):681-689.

Pendlebury D, Wang R, Henin RD, et al. Sequence and conformational specificity in substrate recognition: several human Kunitz protease inhibitor domains are specific substrates of mesotrypsin. J Biol Chem. 2014;289(47):32783-32797.

Ponte P, Gonzalez-DeWhitt P, Schilling J, et al. A new A4 amyloid mRNA contains a domain homologous to serine proteinase inhibitors. Nature. 1988;331(6156):525-527.

Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem J. 2004;378(Pt 3):705-716.

Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

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. Mol Biochem Parasitol. 2010;169(1):1-11.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-D489.

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;15:7.