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

Laia Torres Masdeu[1,*]

[1]Department of Pharmacology and Biotecnology, Alma Mater Studorum – Università di Bologna, Bologna, Italy

[*]Corresponding author. laia.torresmasdeu@studio.unibo.it

Abstract

Motivation: The Kunitz-type protease inhibitor domain is a functionally important protein domain involved in protease regulation. This study aimed to build a Hidden Markov Model (HMM) profile for the Kunitz domain to capture its sequence and structural features, enabling accurate identification and annotation of new domain sequences. The motivation behind this research was to build an HMM profile and use it to annotate Kunitz domains in the SwissProt database.
Results: To ensure the robustness of the model, four different models were constructed using structures from three different databases. The models were evaluated and compared to determine if any bias existed in predicting Kunitz proteins based on the input sequences. Surprisingly, no significant differences were observed among the four models. The results indicate that regardless of the database used, the constructed HMMs were equally effective in predicting Kunitz proteins. This finding suggests that the performance and accuracy of the HMMs were not influenced by the specific database used for training. In conclusion, the HMM profiles built for the Kunitz domain were able to recognise Kunitz domains and showed consistent and unbiased predictions across different input databases.
Availability and Supplementary information: The code and the files generated to complete the project, as well as the supplementary data, are available at `https://github.com/torresmasdeu/kunitz_HMM_project`.

Introduction

One important aspect of protein function is its ability to interact with other molecules, including enzymes. In this context, protein domains that act as inhibitors of proteases play a crucial role in regulating proteolytic activity and maintaining the balance of biological processes (Laskowski Jr and Kato, 1980). The Kunitz-type protease inhibitor domain is a well-known example of such a domain, possessing a unique fold and exhibiting remarkable biological significance (Shigetomi et al., 2010).

The Kunitz domain is composed of approximately 60-80 residues (Expasy, 2019). It was first identified and characterised in the bovine pancreatic trypsin inhibitor (BPTI) protein (Deisenhofer and Steigemann, 1975) and later found in various other proteins across different organisms, including animals, plants and microbes (Rawlings et al., 2004).

Functionally, the significance of Kunitz domains lies in their ability to tightly bind and inhibit serine proteases (Vincent and Lazdunski, 1973), which are enzymes involved in numerous physiological processes whose activity must be controlled. By blocking the active site of the protease or interfering with its catalytic mechanism, Kunitz domains can regulate proteolytic activities and control key biological processes such as blood coagulation, fibrinolysis, and immune responses, as well as neuronal activity modulation, embryonic development, and defense against microbial pathogens (Shigetomi et al., 2010; Ranasinghe and McManus, 2013; Wan et al., 2013; Báez et al., 2015). For instance, in blood coagulation pathways, Kunitz domains inhibit key proteases involved in clot formation, maintaining the balance between clotting and anticoagulation, and reducing bleeding (Masci et al., 2000).

Kunitz domains are characterized by their conserved structural motif, which forms a compact and stable fold stabilized by disulfide bridges. Specifically, it is constrained by three disulfide bonds, adding up to six cysteines (InterPro, 2023). As part of the BPTI-like superfamily, it is characterised by a disulfide-rich alpha+beta fold (SCOP, 2021), which consists of a two-stranded antiparallel beta-sheet followed by an alpha-helix (Figure 1). The beta-sheet forms a loop structure containing several conserved residues critical for protease inhibition. The unique arrangement of these secondary structural elements contributes to the inhibitory function of Kunitz domains.

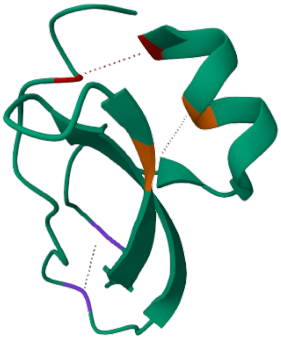

Fig. 1: Kunitz domain of PDB entry 1AAP (PDB, 1991). Coloured residues and bonds correspond to the conserved cysteines and disulfide bonds, respectively.

Given the biological significance of Kunitz domains and their structural conservation, they represent an excellent opportunity for building profile Hidden Markov Models (HMMs) to capture their sequence and structural features. HMMs are statistical models widely used in bioinformatics for representing and analyzing protein families and domains. Constructing an HMM for the Kunitz-type protease inhibitor domain can effectively profile its sequence patterns, detect homologous domains across different organisms, and gain insights into the evolutionary relationships among these domains. Moreover, the availability of extensive sequence databases and advanced computational tools enables the efficient gathering of Kunitz domain sequences from various organisms. By harnessing these resources, comprehensive and representative datasets can be constructed for building the profile HMM. Such a model would allow us to accurately identify and annotate new sequences of the Kunitz domain, predict their structures, and infer functional characteristics.

In summary, the Kunitz-type protease inhibitor domain, with its conserved fold and essential role in protease regulation, provides a compelling case for building a profile HMM. This project has two main aims: to build a model for the Kunitz-type domains, starting from available structural information, and to then use the model to annotate Kunitz domains in SwissProt.

Materials and Methods

Database and Software Tools

Database retrieval

1. UniProt/SwissProt (UniProt, 2023)
   - UniProt database (Release 2023_02) was accessed to retrieve protein sequences.
   - Protein identifiers and sequences were obtained to generate the training, cross-validation, testing and validation tests.
2. Protein Data Bank (PDB) (PDB, 2023)
   - PDB was used on 27/05/2023 to retrieve the structural sequences of the proteins used to make the model.
   - Protein structures were retrieved using PDB Advanced Search, following the parameters detailed in the Structure selection section.

Software Tools

1. PDBeFold (EMBL-EBI, 2014)
   - PDBeFold (version 2.59) was utilised to construct the Multiple Sequence Alignment (MSA) of the selected structures.
   - Structural alignments in FASTA format and RMSD scores were calculated using the PDBeFold web server.
2. UniProt
   - UniProt ID Mapping and Align tools were used to process some of the identifiers and alignments of the models, respectively
   - ID Mapping was used to obtain the UniProtIDs of the proteins used in the training set, while Align was used to check for duplicated sequences.
3. HMMER (Hidden Markov Model) (HMMER, 2023)
   - HMMER (version 3.3.2 - Nov 2020) software package was employed for profile Hidden Markov Model (HMM) building and analysis.
   - HMM profiles were constructed using the HMMER software suite (specifically, `hmmbuild` and `hmmsearch`).

HMM Construction and Validation

For more details on the computing methods, including the code used and files generated, refer to the Supplementary Materials section.

Structure selection

The structures used to build the Hidden Markov Model (HMM) profile from PDB were retrieved using the following Advanced Search:

- Kunitz domain identifier: three different classification codes from three different classification databases were used to create four different models. The identifiers and procedures used are detailed in the Comparative Analysis of Multiple HMM Models section.
- Resolution: smaller than 3Å.
- Entity sequence length: between 51 and 76 residues (Expasy, 2019).
- Mutation count: 0.

These searches retrieved 112 (M1), 144 (M2) and 13 (M3) protein structures. However, as structures belonging to the same protein can be found in different PDB files, the polymer entities (chains matching the query) were grouped by 95% sequence identity (to exclude any possible biases caused by the experimental design). This procedure returned 18 (M1), 20 (M2) and 13 (M3) representatives (Table 1).

Moreover, a manual inspection was made to discard any additional sequences that might contain non-annotated mutations (by checking the title and articles associated with the entries).

HMM generation

Once the structures were selected, they were submitted to PDBeFold, which retrieved their MSA. The output was analysed to check the accuracy of the alignment, and those proteins that had an RMSD higher than 2Å were discarded. Lastly, the MSA was trimmed to remove the poorly aligned/highly gapped regions, as

Table 1. Comparison between retrieved and final representatives

	PDB output	HMM profile input
Model 1 (Pfam)	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1YC0:I,1ZR0:B, 3BYB:A,3M7Q:B,4DTG:K, 4ISO:B,4NTW:B,4U30:X, 4U32:X,5M4V:A,5PTI:A, 5YV7:A,6Q61:A,6YHY:A	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1ZR0:B,3BYB:A, 3M7Q:B,4DTG:K,4ISO:B, 4NTW:B,4U30:X,4U32:X, 5M4V:A,5PTI:A,5YV7:A, 6Q61:A,6YHY:A
Model 2 (CATH)	1AAP:A,1BUN:B,1D0D:A, 1DTX:A,1KTH:A,1YC0:I, 1ZR0:B,2UUY:B,2W8X:A, 3BYB:A,3M7Q:B,4DTG:K, 4ISO:B,4NTW:B,4U30:X, 4U32:X,5M4V:A,5PTI:A, 5YV7:A,6Q61:A	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1ZR0:B,3BYB:A, 3M7Q:B,4DTG:K,4ISO:B, 4NTW:B,4U30:X,4U32:X, 5M4V:A,5PTI:A,5YV7:A, 6Q61:A
Model 3 (SCOP)	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1TFX:C,1Y62:A, 1ZR0:B,3BYB:A,4ISL:B, 4NTW:B,4U32:X,5M4V:A, 5PTI:A	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1TFX:C,1Y62:A, 1ZR0:B, 3BYB:A,4ISL:B, 4NTW:B,4U32:X,5M4V:A, 5PTI:A
Model 4 (PCS)	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1TFX:C,1Y62:A, 1YC0:I,1ZR0:B,3BYB:A, 3M7Q:B,4DTG:K,4ISL:B, 4ISO:B,4NTW:B,4U30:X, 4U32:X,5M4V:A,5PTI:A, 5YV7:A,6Q61:A,6YHY:A	1AAP:A,1BUN:B,1DTX:A, 1KTH:A,1Y62:A,1ZR0:B, 3BYB:A,3M7Q:B,4DTG:K, 4ISO:B,4NTW:B,4U30:X, 4U32:X,5M4V:A,5PTI:A, 5YV7:A,6Q61:A,6YHY:A

On the left column, representatives obtained after the PDB Advanced Search. On the right, final representatives resulting from RMSD-MSA value verification, trimming and duplicate removal.

[1] In red, removed PDB representatives.

[2] In teal, final representatives found in all databases.

[3] In orange, final representatives belonging just to the Pfam database.

[4] In violet, final representatives belonging just to the SCOP database.

these could have a negative influence on the resulting model. The resulting file was then uploaded to the Align tool of the UniProt website, and the duplicated sequences were removed. This resulted in obtaining, after all of this preprocessing, 17 (M1), 16 (M2), 13 (M3) and 18 (M4) sequences to build each model (Table 1).

The Hidden Markov Model profiles were generated by running the `hmmbuild` program of HMMER with default options, taking as input the trimmed FASTA files. To verify that the trained HMM was able to correctly recognise the proteins of the dataset, its performance was checked by running the `hmmsearch` program of HMMER, with the `--max` option (which turns off all the heuristics for cutting off distantly related proteins) over the training set, composed of the FASTA sequences of the structures used to train the model.

Method testing

Validation set

Once the model was determined as consistent with regards to the training set, its efficiency to annotate Kunitz domains in SwissProt had to be tested. To this end, a validating set was created, consisting of two subsets: positive –sequences with an annotated Kunitz domain– and negative set –proteins lacking a Kunitz domain. The positive set was retrieved through the Advanced Search of the UniProt database, by querying those proteins that had been annotated with the Pfam Kunitz domain identifier (PF00014), and by selecting only those entries that were found in SwissProt (manually curated). The negative set contained the remaining sequences in SwissProt. This procedure returned 390 proteins for the positive set and 569126 for the negative set. To avoid biases, those sequences that were used to make the HMM profile were removed from the positive set. These resulted in 374 (M1, M2, M4) and 377 (M3) final proteins in the positive set. These two subsets were then unified into one validating set and were run against the model (using the `hmmsearch` program of HMMER, with the `--max` and `--noali` options, with the latter excluding the alignments from the output, to simplify its visualisation). The hits of this procedure were refined –by selecting their UniProtID and best 1 domain E-value–, and classified whether they belonged to the positive (1) or negative (0) class. Those proteins belonging to the negative set that had not been matched to the model were assigned an arbitrarily high E-value (100). With this, the data was ready to be used for the model evaluation.

Cross-validation and testing sets

The testing procedure consisted of the implementation of a 2-fold cross-validation test: the positive and negative sets were split into two subsets, so as to optimise the classification (E-value) threshold on one subset (cross-validation set) to test its performance on the other subset (testing set), and reversibly.

Model evaluation

The evaluation of the model was conducted by computing the scoring indexes on the validation set, using a Python script (Supplementary Materials section). It was conducted in 5 sub-steps:

1. Calculating the optimised E-value on the first dataset
2. Using this E-value on the second dataset and evaluating its performances
3. Calculating the optimised E-value on the second dataset
4. Using this E-value on the first dataset and evaluating its performances
5. Computing the average E-value and getting the performances on the whole dataset

Specifically, the scoring indexes that the script returned were the Matthews correlation coefficient (MCC), accuracy score (ACC) and a confusion matrix (CM), which showed the distribution of the true positive (TP) true negative (TN), false positive (FP) and false negative (FN) entries.

Comparative Analysis of Multiple HMM Models

The proteins belonging to the positive set were retrieved by querying in UniProtKB for those reviewed sequences with a Kunitz domain with the Pfam identifier, UniProt Advanced Search accepts search with neither CATH nor SCOP2 identifiers.

Therefore, using only the Pfam identifier to retrieve the PDB structures that constituted the model could potentially lead to a biased model. Thus, 4 model-building approaches were developed:

1. Pfam-based model (M1): PDB database was queried to fetch those structures annotated with the Kunitz domain in the Pfam database (Pfam Identifier PF00014)
2. CATH-based model (M2): PDB database was queried to fetch those structures annotated with the Kunitz domain in the CATH database (Lineage Identifier 4.10.410.10)
3. SCOP-based model (M3): PDB database was queried to fetch those structures annotated with the Kunitz domain in the SCOP2 database (Lineage Identifier 4003337)
4. Pfam+CATH+SCOP-based model (M4): PDB entry IDs for all of the structures retrieved in the previous three models were unified and uniquely sorted (to refrain from containing duplicated entries).

Each model was independently constructed and evaluated.

Results and discussion

Structure selection and data preprocessing

During the PDB Advanced Search, to find the structures that matched the query for each of the 4 models, Mutation Count = 0 was added to not include those proteins that had been willingly mutated for experimental issues, as these could potentially have an impact on the resulting model, which could contain positions that do not belong to the actual wild-type domain. However, this did not remove all of the mutated entries, as some did not have the mutations computed in the count, and therefore a manual inspection of every representative entry was conducted. These representatives were obtained by grouping the single sequences with a 95% sequence similarity, which could be there if you had a manually-engineered mutation or not. However, this 5% of dissimilarity could also be given by the experimental method through which this sequence was obtained. Therefore, using a 100% sequence identity threshold, as was initially thought of, would be incorrect, because the grouped sequences could potentially be biased by the fact that different entries for the same protein were obtained through different methods.

Once this step was performed, each group of structures for each model were aligned, and the resulting MSAs was inspected. For model 2 (CATH-based) 3 of the entries (1D0D, 2UUY and 2W8X) displayed a high RMSD (<2Å) (Supplementary Figure 2, in the Supplementary Materials section). All three of them belonged to tick (*Rhipicephalus appendiculatus*) proteins which, according to the literature associated with the PDB entries (Paesen et al., 2009, 2007; Charles et al., 2000), had a domain resembling Kunitz proteins, but with notable conformational changes on the classical domain, which may explain why, despite being classified as such, did not have a low RMSD value in comparison with the other retrieved structures. Interestingly, neither Pfam nor SCOP2 databases annotated these entries as Kunitz domains, which could explain that the differences were not supported by their sorting models.

Lastly, the correct MSAs were trimmed to remove any starting or ending highly gapped regions. After this step, the trimmed MSAs were run through UniProt's Align tool to check that no entries were duplicated: for model 1 (Pfam-based) and model 2 (CATH-based), 1YC0:I entry was removed, as it was identical to 4ISO:B. For model 4 (Pfam+CATH+SCOP-based), there were two different identical clusters: 1YC0:I, 4ISL:B and 4ISO:B, and 1TFX:C and 4DTG:K. In both cases, the latter entry was kept.

Training sets

Once the MSAs were supervised, the models were built (Figure 2). To check their consistency, each model was run against its corresponding training set. All yielded very good results, with good sequence coverage, consistency and alignment quality, as well as very low E-values (all sequences for every model had a best 1 domain E-value smaller than `1.3e-18`).

Both for the training and validation sets, the E-values that were selected were those of the best 1 domain, the best local alignment, as the aim was to see if, inside each particular protein sequence, one could find the domain described by the model.

Validation sets

The validation set was divided into two subsets: the positive (those proteins that were annotated with the Pfam Kunitz domain) and the negative sets, with entries belonging to the manually curated SwissProt database. The positive set for each model was later reduced so as to not contain the UniProt entries of the PDB entities used to make the model. As each model had different proteins as input, the size of the positive set changed accordingly. It is worth mentioning that the number of removed sequences from the positive set did not correlate directly to the number of PDB

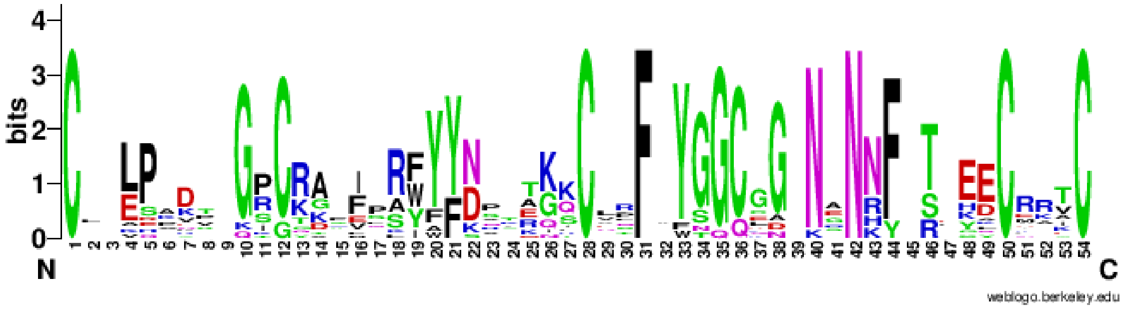

Fig. 2: Sequence Logo of M1 (Pfam). This graphical representation shows the conservation of amino acids at each position in the sequence alignment. Observe how the six cysteines are highly conserved. To access the other sequence logos, go to Supplementary Figure 1, in the Supplementary Materials section. Obtained from Crooks et al. (2004)

Table 2. Evaluation results for the four models

	Model 1			Model 2			Model 3			Model 4		
Set type	1	2	Validation	1	2	Validation	1	2	Validation	1	2	Validation
CV E-val	`1e-5`	`1e-3`	–	`1e-3`	`1e-5`	–	`1e-5`	`1e-3`	–	`1e-4`	`1e-3`	–
Average E-val	–	–	`5.05e-4`	–	–	`5.05e-4`	–	–	`5.05e-4`	–	–	`5.5e-4`
ACC	`0.9999`	`1`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`	`0.9999`
MCC	`0.9919`	`1`	`0.9973`	`0.9973`	`0.9973`	`0.9973`	`0.9946`	`0.9973`	`0.9960`	`0.9946`	`0.9973`	`0.9973`
FN	3	0	2	1	1	2	2	1	3	2	1	2

Model evaluation was done by applying a 2-fold cross-validation set: taking one set (cross-validation set), finding its best E-value, and applying this optimised E-value to the other set (testing set). This was done for both sets. Finally, the average between the two best E-values was applied to the total validation set.

CV: Cross-validation set, ACC: accuracy score, MCC: Matthews Correlation Coefficient, FN: False negative values

representatives used to build the model. This was because either some PDB representatives shared the same UniProtID, or because they had no correspondent UniProtID.

To validate the model, both subsets were unified into a unique validation set, which explains why the `-Z` option of the `hmmsearch` command was not included in this report. This parameter normalises the space search, by dividing the E-value by the number of sequences provided in input. As both subsets were run together, there was no need to normalise any data; had the positive set and negative set been searched against the model separately, this should have been indispensable, as the size of the two subsets would have been significantly different (≈370 vs. 569126, respectively).

Model evaluation

The four models were developed after finding out that UniProt only allowed querying the Kunitz domain through the Pfam identifier. There could be a potential bias in using only PDB structures found within the Pfam database (M1), and as such, it was decided to do 2 more models, with structures labelled with lineage identifiers from two other known and used databases, CATH (M2) and SCOP2 (M3), and a final fourth model that combined the three individual models.

After implementing the 2-fold cross-validation and test sets, both model 1 and model 2 performed at their best with an E-value of `5.05e-4`, with an accuracy score (ACC) of 0.999 and an MCC of 0.997, and 2 false negative results (Table 2). It is important to note, both for these models and the other two, that the mentioned E-values are approximate; they are not the exact value at which the metrics change; however, the exact E-value is in the same order of magnitude (Supplementary Materials). This insignificant difference between the optimal E-value of M1 and M2 can be explained because the structure dataset used for the model, after all the preprocessing steps, only differed in one PDB entry (Table 1). Therefore, the model which was obtained from structures only labelled with a CATH lineage identifier was equally good at predicting the sequences on the positive set as Pfam, which could indicate that the model used by Pfam and CATH is similar, as they are both capable of identifying the same protein domains.

Model 3 performed at its best also with an E-value of `5.05e-4`, but with a bit lower scoring indexes: an ACC of 0.999 and an MCC of 0.996, and 3 false negative results (Table 2). To get the same ACC, MCC and FN values as the ones of model 1 and model 2, the E-value had to be set at `6e-4`. Despite the changes being considerably insignificant, as seen in Table 1, there were 10 different sequences between the Pfam and SCOP input structures, which could explain the slight E-value change between the models.

The fourth model, which was built with the sequences annotated by any of the 3 previously-modelled databases, yielded very similar results to the previous models: at an E-value of `5.05e-5`, it only found 2 false negatives and no false positives (Table 3), with an accuracy score of 0.999 and an MCC of 0.997 (Table 2).

Table 3. Confusion matrix for model 4

		Real	
		Positive	Negative
Predicted	Positive	752	0
	Negative	2	569126

From top left to bottom right: True positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

Therefore, these results demonstrate that, whatever dataset (or combination of datasets) you use, there are no statistical differences between the performance of the model and its ability to predict the presence of a Kunitz domain(s) in a given SwissProt protein.

As mentioned, all models, at their best E-value, were unable to predict correctly the presence of a Kunitz domain in two proteins, with O62247 and D3GGZ8 UniProtIDs. When inspected on the UniProt website, both entries, which belong to the bli-5 protein of *Caenorhabditis elegans*, had a "Caution" remark that stated that they appeared to have serine protease activity *in vitro*, but that it was uncertain if this activity was genuine, as the protein lacked the catalytic features of serine proteases. This led to the careful inspection of the alignment performed by `hmmsearch`, which further supported this theory: taking as an example M4, 36 and 39 residues of the proteins (respectively) had been aligned (to some extent) to the model, out of 51 positions that it had. Moreover, among these partially aligned residues, 2 of the 6 C that are crucial for the Kunitz domain folding, were not present in the protein sequences Supplementary Figure 3, in the Supplementary Materials section. Altogether, these findings could support the fact that these proteins were badly annotated.

Conclusions

This project had two aims, to build a model for the Kunitz domain from structural information, and to use this model to search and annotate Kunitz domains in SwissProt. As per the results, all four models were able to predict with high accuracy, while generating a negligible number of false positives and false negatives. The FN that were obtained could be due to annotation errors. Moreover, no significant differences were seen when evaluating model 1 with respect to the other models. This allows us to conclude that using the Pfam database to build the HMM profile and look for Kunitz domains within proteins did not bias the results.

Supplementary Materials

To access the supplementary materials, please go to the following url: https://github.com/torresmasdeu/kunitz_HMM_project. Supplementary images are located inside the directory with the same name.

Bibliography

A. Báez, E. Salceda, M. Fló, M. Grana, C. Fernández, R. Vega, and E. Soto. α-dendrotoxin inhibits the asic current in dorsal root ganglion neurons from rat. *Neuroscience letters*, 606:42–47, 2015.

R. S. Charles, K. Padmanabhan, R. Arni, K. Padmanabhan, and A. Tulinsky. Structure of tick anticoagulant peptide at 1.6 å resolution complexed with bovine pancreatic trypsin inhibitor. *Protein Science*, 9(2):265–272, 2000.

G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6): 1188–1190, 2004.

J. Deisenhofer and W. Steigemann. Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1. 5 å resolution. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry*, 31(1):238–250, 1975.

EMBL-EBI. Pdbe fold v2.59, 2014. URL https://www.ebi.ac.uk/msd-srv/ssm/.

Expasy. Pru00031, Nov 2019. URL https://prosite.expasy.org/rule/PRU00031.

HMMER. Hmmer v3.3.2, 2023. URL http://hmmer.org/.

InterPro. Pancreatic trypsin inhibitor kunitz domain, 2023. URL https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002223/.

M. Laskowski Jr and I. Kato. Protein inhibitors of proteinases. *Annual review of biochemistry*, 49(1):593–626, 1980.

P. Masci, A. Whitaker, L. Sparrow, J. De Jersey, D. Winzor, D. Watters, M. Lavin, and P. Gaffney. Textilinins from pseudonaja textilis textilis. characterization of two plasmin inhibitors that reduce bleeding in an animal model. *Blood coagulation & fibrinolysis*, 11(4):385–393, 2000.

G. C. Paesen, C. Siebold, K. Harlos, M. F. Peacey, P. A. Nuttall, and D. I. Stuart. A tick protein with a modified kunitz fold inhibits human tryptase. *Journal of molecular biology*, 368(4): 1172–1186, 2007.

G. C. Paesen, C. Siebold, M. L. Dallas, C. Peers, K. Harlos, P. A. Nuttall, M. A. Nunn, D. I. Stuart, and R. M. Esnouf. An ion-channel modulator from the saliva of the brown ear tick has a highly modified kunitz/bpti structure. *Journal of molecular biology*, 389(4):734–747, 2009.

PDB. 1aap, 1991. URL https://www.rcsb.org/structure/1AAP/.

R. PDB. Rcsb pdb, 2023. URL https://www.rcsb.org/.

S. Ranasinghe and D. P. McManus. Structure and function of invertebrate kunitz serine protease inhibitors. *Developmental & Comparative Immunology*, 39(3):219–227, 2013.

N. D. Rawlings, D. P. Tolle, and A. J. Barrett. Evolutionary families of peptidase inhibitors. *Biochemical Journal*, 378(3): 705–716, 2004.

SCOP. Bpti-like fold, 2021. URL https://scop.mrc-lmb.cam.ac.uk/term/2000414.

H. Shigetomi, A. Onogi, H. Kajiwara, S. Yoshida, N. Furukawa, S. Haruta, Y. Tanase, S. Kanayama, T. Noguchi, Y. Yamada, et al. Anti-inflammatory actions of serine protease inhibitors containing the kunitz domain. *Inflammation research*, 59:679–687, 2010.

UniProt. Uniprot release 2023_02, 2023. URL https://www.uniprot.org/.

J.-P. Vincent and M. Lazdunski. The interaction between α-chymotrypsin and pancreatic trypsin inhibitor (kunitz inhibitor) kinetic and thermodynamic properties. *European Journal of Biochemistry*, 38(2):365–372, 1973.

H. Wan, K. S. Lee, B. Y. Kim, F. M. Zou, H. J. Yoon, Y. H. Je, J. Li, and B. R. Jin. A spider-derived kunitz-type serine protease inhibitor that acts as a plasmin inhibitor and an elastase inhibitor. *PLoS One*, 8(1):e53343, 2013.