Laboratory of Bioinformatics

20th May 2024

Project report

HMM profile for the Kunitz domain based on a structural alignment

Laura Claudia Verdesca

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

Abstract

Peptidases are crucial for organismal survival, breaking down proteins, but necessitate precise control, they regulate protease activity tightly and serve as switches in signaling pathways. Eukaryotic proteases, like serine, cysteine, and aspartic proteases, are categorized based on sequence homology, reactive site position, and structural characteristics. Inhibitors, including the Kunitz-type family, interact with enzymes in a substrate-like manner. Kunitz-type inhibitors, found across kingdoms, adopt a conserved fold stabilized by disulfide bridges. Examples, like aprotinin (bovine pancreatic trypsin inhibitor, BPTI), serve diverse functions. Structurally, it features a stable fold with disulfide bonds and positively-charged side chains, conferring stability and inhibitory properties. Hidden Markov Models (HMMs) aid in protein domain identification by capturing sequence patterns. This study successfully constructed a Hidden Markov Model (HMM) for the Kunitz BPTI domain using a dataset of structurally characterized proteins. The model effectively identified the domain in new seed sequences, demonstrating high accuracy at e-value thresholds between 1e-07 and 1e-08 and minimal false positives and negatives, making the model suitable for functional annotation of unreviewed proteins.

Keywords: Protease inhibitors, Kunitz domain, Hidden Markov Model, Cross validation

Supplementary material: https://github.com/LauraVerdesca/LB1_report

Introduction

Peptidases are essential for the survival of all types of organisms since they break down proteins, but their activty must be carefully controlled. Protease inhibitors play critical roles in natural systems by tightly regulating protease activity and serving as switches in many signaling pathways.

Eukaryotic proteases, including serine, cysteine, and aspartic proteases, can be categorized into different families based on sequence homology, reactive site position, structural characteristics, and mechanism of action. These inhibitors encompass various families, such as the Kunitz-type family, which binds in an extended β-sheet with the enzyme in a substrate-like manner. Kunitz-type serine protease inhibitors are present in animals, plants, and microbes. They typically consist of 50–70 amino acids and adopt a conserved structural fold with antiparallel β-sheets and helical regions stabilized by disulfide bridges.

Examples of Kunitz-type protease inhibitors include: the Kunitz-type toxin in venomous animals like snakes, spiders, and scorpions, aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI). [1] [2]

In this project we are going to focus on the bovine pancreatic trypsin inhibitor (BPTI) or aprotinin. Aprotinin, marketed under the brand name Trasylol by Bayer and later by Nordic Group Pharmaceuticals, is the small protein bovine pancreatic trypsin inhibitor (BPTI). This molecule acts as an antifibrinolytic agent, effectively inhibiting trypsin and related proteolytic enzymes. It is administered via injection, typically during complex surgeries like those involving the heart or liver, Trasylol works by slowing down fibrinolysis, the natural process of breaking down blood clots. Its primary goal is to reduce bleeding during surgery, hence minimizing the need for blood transfusions and preventing end-organ damage caused by hypotension resulting from excessive blood loss.

Physiological functions of BPTI include the protective inhibition of the major digestive enzyme trypsin when small amounts are produced by cleavage of the trypsinogen precursor during storage in the pancreas.

Bovine pancreatic trypsin inhibitor is an extensively studied model structure. The majority are restricted to metazoa with a single exception: Amsacta moorei entomopoxvirus, a species of poxvirus. The structure is a disulfide rich alpha+beta fold.

Aprotinin, derived from bovine lung tissue, is a monomeric globular polypeptide characterized by a molecular weight of 6.5 kDa. Comprised of a 58-residue chain, it adopts a stable and compact tertiary structure categorized as 'small SS-rich' type. This structure features three disulfide bonds, a twisted β-hairpin, and a C-terminal α-helix.

Aprotinin contains 10 positively-charged lysine (K) and arginine (R) side chains and only 4 negative aspartate (D) and glutamates (E), confering a basic nature to the protein.

The presence of 3 disulfide bonds linking the 6 cysteine members of the chain (Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51) makes the protein extremely stable. Furthermore, its inhibitory mechanism against trypsin involves the tight binding of the exposed loop's long lysine 15 side chain to the specificity pocket at the enzyme's active site, thereby impeding its enzymatic activity. Initially synthesized as a longer precursor sequence, BPTI undergoes folding before being cleaved into its mature form. [4] [5]

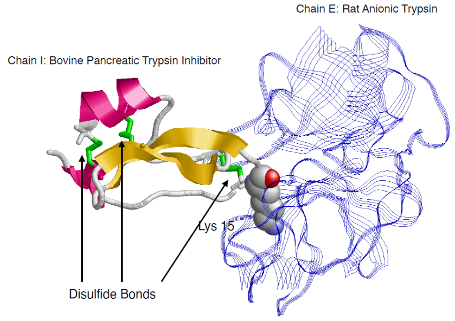
Chain E: Rat Anionic Trypsin
Chain I: Bovine Pancreatic Trypsin Inhibitor
Lys 15
Disulfide Bonds

Figure 1- This is one of the possible folding of the BPTI protein.

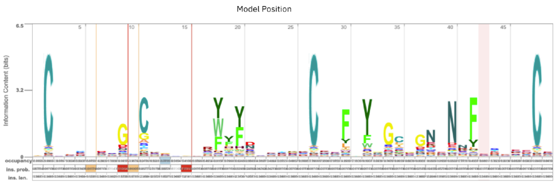

Figure 2 - From Interpro: Conservation of cysteins, expected since they mainly contribute to the stability of the protein, but we do have also other conserved amino acids.

Hidden Markov Models (HMMs) are a powerful tool for protein domain identification, they can be used to represent and analyze the sequence patterns and evolutionary relationships within protein domains. These profiles capture the statistical characteristics of amino acid sequences associated with specific protein domains, allowing for the identification of homologous domains in newly sequenced proteins. HMMs are trained on a diverse set of known protein domain sequences, learning the sequence patterns, conservation, and variability within these domains. Once trained, HMMs can effectively search large databases of protein sequences to identify matches to known domains, even in cases of low sequence similarity. This approach is particularly useful for annotating protein sequences, predicting domain architectures, and inferring protein function based on domain composition. [6]

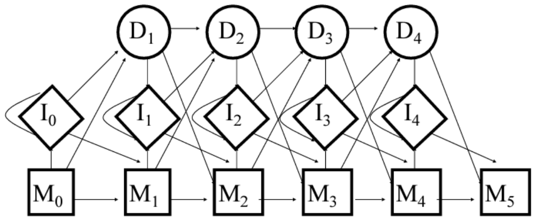

Figure 3 - Schematic view of an HMM model

Aim

The main aim of this project is to build a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain. The model will be developed from available structural information, the reason behind this choice is that structure offers a more comprehensive understanding of how proteins fold, interact, and function in biological systems, while the simple naked sequence of protein gives us a narrower view of the protein.

Then we want to use the model to annotate the Kunitz domain in Swiss Prot.

Methods

Data collection - 3D structures

The dataset to build the HMM profile was assembled following these steps:

1. Advanced search on RCSB PDB [7] using filters: Pfam ID PF00014; resolution below 3˚A; sequence length between 50 to 80 amino residues and mutation count equal to 0. 131 sequences were retrieved.
2. The sequences were clustered using the command `blastclust` [8] based on a sequence identity threshold of 80% to select representatives, thereby avoiding potential bias in the model construction from redundant structures. This process yielded a total of 20 sequences for analysis. (rcsb_pdb_custom_report_20240414163443.csv)

3D Alignment

The dataset was loaded in PDBe-Fold [9] to perform the multiple structure alignment with default parameters: the overall RMSD was 0.75˚A. The alignment has been used to build the model. It appeared that the first 10 amino acids of the alignment contained mostly gaps, therefore they were selected as candidates for trimming, which was performed in the building step of the model.
The Weblogo tool [10] was employed to showcase the sequence profile of the Kunitz domain. (clean_kunitz_3d.aln)

Generation of Hidden Markov Model

HMMER 3.4 (August 2023) [11] software package was used to generate a hidden Markov model for the Kunitz domain, `hmbuild` command with default parameters was used to construct a HMM profile of the Kunitz domain. The model was generated using the trimmed version of the alignment. (clean_kunits_3d.hmm)

Generation of the testing dataset

To evaluate and test the model performance, we obtained two sets of sequences, one ‘positive’ set with sequences containing the BPTI-Kunitz domain and one ‘negative’ set containing sequences lacking the Kunitz domain.
The positive dataset was generated by using the Advanced Search tool in Uniprot [12], the query was designed to retrieve proteins containing a cross-reference to Pfam entry PF00014 which are also reviewed, but which do not have a cross-reference to PDB to avoid using the same proteins used for the model. This dataset contains 358 proteins. Since the positive dataset will be used for testing, we need to remove any sequence with a high degree of similarity to the sequence present in the dataset used for training, to achieve this we employed the `blastp` command and removed each sequence with a 95% similarity. (bpti_pos_selected.fasta)
The negative dataset was generated by using the Advanced Search tool in Uniprot, the query was designed to retrieve proteins which are reviewed, that lack a cross-reference to Pfam entry PF00014 and have a sequence lenght over 40 aminoacids. This dataset contains 570891 proteins. (swiss_negatives.fasta)

Testing, training and model optimization

To evaluate the performance and generalization ability of our machine learning model we employ two types of cross-validation techniques: 2-fold and 5-fold. The technique involves dividing the dataset into complementary subsets, typically training and validation sets, and iteratively training the model on one subset while evaluating its performance on the other. This process is repeated multiple times, with different partitions, and performance metrics are averaged across all iterations.

2-fold cross validation

For validation we used HMMER 3.4 software package and applied a two-fold cross validation technique. Initially, we randomly shuffled the dataset and divided it into two equal-sized sets, denoted as set-1 and set-2. Subsequently, we trained the model on set-1 and evaluated its performance on set-2, followed by training on set-2 and validating on set-1. The decision to utilize two folds was influenced by the lengths of the datasets; further partitioning would have resulted in lists where positive instances were notably underrepresented compared to negatives.

5-fold cross validation

We later implemented five-fold cross validation, as before we randomly shuffled the dataset and divided it into five equal-sized parts. Later we trained the model on four out of five sets and evaluated on the fifth remaining one, this operation was repeated iteratively until both training and evaluated were performed once on every subset and combination of them. (cross_val.sh) We decided to also use five-fold cross validation as it averages performance metrics over five different measures rather than two, providing more robust statistics of the model.

In both cross validation techniques we used a python script to compute the performances of the model, the metrics we chose are: accuracy (ACC), Matthews Correlation Coefficient (MCC) and F1 score. (best_mcc.py, Performance.py)

- Accuracy (ACC) is a commonly used performance metric that measures the proportion of correctly classified instances out of the total number of instances in a dataset. It is calculated as the ratio of the number of correctly predicted instances to the total number of instances.

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

- Matthews Correlation Coefficient (MCC) is a performance metric that takes into account true positives, true negatives, false positives, and false negatives to provide a balanced measure of classification performance, particularly significant for imbalanced datasets. It ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between predictions and observations.

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

- The F1 score is a key metric for evaluating binary classification models, especially with imbalanced datasets. It is the harmonic mean of precision and recall, balancing the accuracy of positive predictions (precision) with the model's ability to identify all positive instances (recall).

$$\text{Precision (P)} = \frac{TP}{TP+FP}$$

$$\text{Recall (R)} = \frac{TP}{TP+FN}$$

$$\text{F1} = 2 \times \frac{P\times R}{P+R}$$

The script processes a file containing proteins alongside their respective e-values and class assignments. By specifying a threshold, it calculates essential metrics including the confusion matrix, accuracy, Matthews Correlation Coefficient (MCC) and F1 score.
The negative class greatly outbumbers the positive one, leading to a skewed class distribution, therefore we decide to prioritize MCC and F1 score over accuracy, as they are a better performance metric.
The performance was calculated with E-value thresholds in range of 10^{-1} to 10^{-15}, with a step of 10^{-1} between each tested E-value. The threshold which yielded the best performance on the first subset, according to MCC, was chosen and tested on the opposite subset. The mean of the E-values was chosen as the overall threshold. Finally, the performance of the model on the entire dataset was measured as well using the overall threshold. We adopted the same logic for five-fold cross validation.

Results

The first step in our study is to create a dataset of proteins displaying the Kunitz domain, to retrieve the structures we perform an advanced search on the PDB website, selecting structures labeled with the PFAM identifier PF00014. We restrict our search to protein structure with high resolution (≤3), a length between 50 and 80 amino acids and a polymer entity mutation count of 0.
We obtained 131 structures, and one possible issue is redundancy. Among these 131 proteins, many of them are probably, if not equal, very similar. To overcome this problem we use `blastclust`, which is a command-line tool used in bioinformatics for clustering sequences based on their pairwise sequence similarity.

We group these structures according to a given threshold of similarity, which we fix at 80% similarity. What we obtained are 20 clusters, for each one of them a single representative was chosen.
We later proceed with the 3D structure alignment of the protein dataset that we just created, the resulting alignment will be used for the generation of the HMM model. The structure alignment produces good values of RMSD, meaning that the level of similarity between the candidate proteins is high.
We computed the sequence logo with Web logo, the picture shows the presence of 6 highly conserved cysteine residues. This is consistent with crystallographic studies which show the presence of three stabilizing disulfide bridges which maintain the structural integrity of the protein [13].

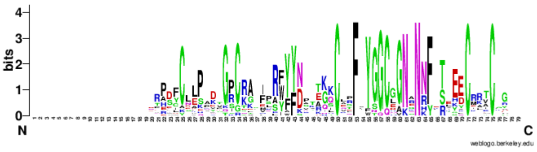

Figure 4 - Sequence logo

We later cleaned the file with `hmmbuild` to discard positions that were not properly aligned, as a result, we obtained proteins with similar lengths aligned together. With `hmmbuild` we create the Hidden Markov model for the Kunitz domain.
To train and test the model, two datasets were built both consisting of Uniprot/Swisspro sequences, a positive set containing 358 reviewed sequences containing the Kunitz-type domain and a negative set, containing 570891 reviewed sequences not containing the Kunitz-type domain.
We later performed testing with 2-cross validation technique and 5-fold cross validation. The model performance was tested by running a script returning the confusion matrix, the accuracy (AC), the Matthew's correlation coefficient (MCC), the TPR (True Positive

Rate) and the FPR (False Positive Rate) values and F1 score for threshold values ranging between 1 and 1e−15.

The ideal threshold selected during two-fold cross validation is equal to 1e-07 and the average MCC value amounted to 0.9958142418530556. More details in the supplementary material. (set_1.res, set_2.res)

In regards of five-fold cross validation we report more precisely the results. We list all the collapsed file combinations results compared to their corresponding final validation result. We plotted for each combined collapsed file the MCC value and the different thresholds, for reference we insert only the one corresponding to the file *com_set_0123.res,* the other plots are present in the supplementary material section. We display also the confusion matrices, corresponding to values of true positives, true negatives, false positives and false negatives that we obtain once we applied the ideal threshold on the testing subset. (com_set*.res, out_best)

	Threshold	Accuracy	MCC	F1
Training set	1e-06	0.99998	0.99165	0.995786
	1e-07	0.99999	0.99577	0.997179
	1e-08	0.99999	0.99576	0.995750
	1e-09	0.99999	0.99576	0.995750
Testing set	1e-07	1.0	1.0	1.0

Table 1 - Collapsed file com_set_0123.res compared to validation set 4

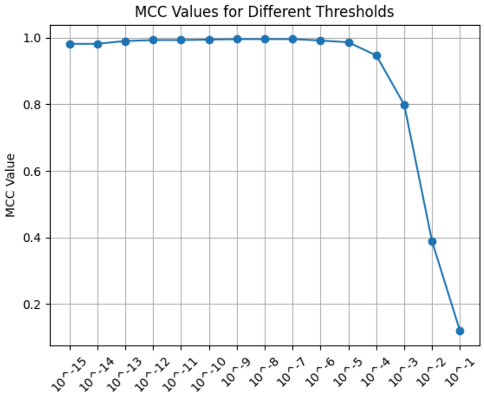

Figure 5 - The plot is produced with matplolib library in Python. On the x axis the threshold values are displayed, while on the y axis we have the MCC value. This plot shows which is the best E-value threshold that produces the highest level of MCC.

	Positives	Negatives
Predicted positives	TP = 71	FP = 0
Predicted negatives	FN = 0	TN = 114179

Table 2 - Confusion matrix of validation set 4 with threshold 1e-7.

	Threshold	Accuracy	MCC	F1
Training set	1e-06	0.99998	0.99165	0.995786
	1e-07	0.99999	0.99577	0.997179
	1e-08	0.99999	0.99576	0.995750
	1e-09	0.99999	0.99576	0.995750
Testing set	1e-07	0.99999	0.99302	0.992805

Table 3 - Collapsed file com_set_0124.res compared to validation set 3

	Positives	Negatives
Predicted positives	TP = 69	FP = 1
Predicted negatives	FN = 0	TN =114208

Table 4 - Confusion matrix of validation set 3 with threshold 1e-7.

	Threshold	Accuracy	MCC	F1
Training set	1e-06	0.99998	0.99165	0.995786
	1e-07	0.99999	0.99577	0.997179
	1e-08	0.99999	0.99576	0.995750
	1e-09	0.99999	0.99576	0.995750
Testing set	1e-07	1.0	1.0	1.0

Table 5 - Collapsed file com_set_0234.res compared to validation set 1

	Positives	Negatives
Predicted positives	TP = 70	FP = 0
Predicted negatives	FN = 0	TN =114142

Table 6 - Confusion matrix of validation set 1 with threshold 1e-7.

	Threshold	Accuracy	MCC	F1
Training set	1e-06	0.99998	0.99165	0.995786
	1e-07	0.99999	0.99577	0.997179
	1e-08	0.99999	0.99576	0.995750
	1e-09	0.99999	0.99576	0.995750
Testing set	1e-07	0.99999	0.99292	0.993103

Table 7 - Collapsed file com_set_1234.res compared to validation set 0

	Positives	Negatives
Predicted positives	TP = 72	FP = 0
Predicted negatives	FN = 1	TN =114188

Table 8 - Confusion matrix of validation set 0 with threshold 1e-7.

	Threshold	Accuracy	MCC	F1
Training set	1e-06	0.99998	0.99165	0.995786
	1e-07	0.99999	0.99577	0.997179
	1e-08	0.99999	0.99576	0.995750
	1e-09	0.99999	0.99576	0.995750
Testing set	1e-07	0.99999	0.99302	0.993006

Table 9 - Collapsed file com_set_0134.res compared to validation set 2

	Positives	Negatives
Predicted positives	TP = 71	FP = 0
Predicted negatives	FN = 1	TN =114174

Table 10 - Confusion matrix of validation set 2 with threshold 1e-7.

The average MCC from the 5-fold CV is 0.995796 and average F1 score is 0.9957832393357077.

Discussion

The sequence logo reveals six highly conserved cysteine residues, corroborating crystallographic studies that show three stabilizing disulfide bridges essential for maintaining the protein's structural integrity. Notably, the first and third disulfide bridges (Cys6-Cys57 and Cys31-Cys53), along with residues Phe34 and Tyr36, exhibit a high level of conservation.

This study aimed to construct a Hidden Markov Model (HMM) specifically for the Kunitz BPTI domain using a dataset of structurally characterized proteins. The model's effectiveness in identifying the domain in new seed sequences was subsequently assessed. Results demonstrate that the HMM-based approach accurately predicts the presence of the domain in known proteins. Optimal performance was observed at E-value thresholds between 1e-07 and 1e-08, where Accuracy, Matthews Correlation Coefficient (MCC), and F1 values were highest, with minimal false positives and false negatives. Thus, this model can be effectively employed for the functional annotation of unreviewed proteins.

Despite the strong performance metrics, some misclassifications were observed, including one false positive protein (D3GGZ8) and two false negatives (P84555 and P56409). Further inspection in Uniprot revealed that P84555 and P56409 are annotated to contain a Kunitz domain within the Family and Domain window but are associated with the InterPro [14] code IPR036880 rather than PFAM 00014. This discrepancy suggests a possible variation within the Kunitz domain, deviating from the canonical structure and leading to misclassification. Additionally, the Kunitz domain in P84555 is unusually short at only 20 amino acids, compared to the typical length of at least 50 amino acids. D3GGZ8 was identified as a false positive, meaning it was incorrectly labeled as containing the Kunitz domain. Examination in Uniprot shows that while this protein is annotated with a Kunitz domain, it has a low annotation score of 2/5, with its structure resolved using Alphafold and displaying numerous disordered regions. When the 3D structure of D3GGZ8 was superimposed with the reference Kunitz domain structure in Chimera [15], only a small portion aligned, highlighting significant discrepancies.

Conclusion

This study successfully constructed a Hidden Markov Model (HMM) for the Kunitz BPTI domain using a dataset of structurally characterized proteins. The model effectively identified the domain in new seed sequences, demonstrating high accuracy at E-value thresholds between 1e-07 and 1e-08 and minimal false positives and negatives, making the model suitable for functional annotation of unreviewed proteins.

However, some misclassifications occurred, specifically one false positive (D3GGZ8) and two false negatives (P84555 and P56409). Further inspection revealed that the false negatives were associated with a different InterPro code (IPR036880) rather than the expected PFAM 00014, suggesting a variation within the Kunitz domain. The false positive, D3GGZ8, although annotated with a Kunitz domain, showed poor structural alignment with the reference Kunitz domain, highlighting potential issues with annotation accuracy. These findings suggest that variations and deviations within the Kunitz domain can lead to misclassification, emphasizing the need for continuous refinement of the model.

References

[1] Mishra, M. Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. *J Mol Evol* **88**, 537–548 (2020). https://doi.org/10.1007/s00239-020-09959-9

[2] de Magalhães, Mariana T Q et al. “Serine protease inhibitors containing a Kunitz domain: their role in modulation of host inflammatory responses and parasite survival.” *Microbes and infection* vol. 20,9-10 (2018): 606-609. doi:10.1016/j.micinf.2018.01.003

[3] Pasternak, A et al. “Comparison of anionic and cationic trypsinogens: the anionic activation domain is more flexible in solution and differs in its mode of BPTI binding in the crystal structure.” *Protein science : a publication of the Protein Society* vol. 8,1 (1999): 253-8. doi:10.1110/ps.8.1.253

[4] Kassell, B, and M Laskowski Sr. “The basic trypsin inhibitor of bovine pancreas. V. The disulfide linkages.” *Biochemical and biophysical research communications* vol. 20,4 (1965): 463-8. doi:10.1016/0006-291x(65)90601-7

[5] Richardson JS (1981). "The anatomy and taxonomy of protein structure". *Advances in Protein Chemistry Volume 34*. Advances in Protein Chemistry. Vol. 34. pp. 167–339. doi:10.1016/S0065-3233(08)60520-3. ISBN 978-0-12-034234-1. PMID 7020376.

[6] B.-J. Yoon. Hidden markov models and their applications in biological sequence analysis. Curr. Genomics, 10, n. 6:402–415, 2009.

[7] Burley, Stephen K., et al. "Protein Data Bank (PDB)." RCSB, 2021, https://www.rcsb.org/.

[8] Altschul, Stephen F., et al. "BLAST." Version 2.11.0, National Center for Biotechnology Information, 1990, https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[9] Krissinel, Evgeny, and Kim Henrick. "PDBeFold." Version 2.56, European Bioinformatics Institute, 2004, http://www.ebi.ac.uk/msd-srv/ssm/.

[10] Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004)

[11] "UniProt: The Universal Protein Knowledgebase in 2023." UniProt, 2023, https://www.uniprot.org/. Accessed 18 May 2024.

[12] "UniProt: The Universal Protein Knowledgebase in 2023." UniProt, 2023, https://www.uniprot.org/. Accessed 18 May 2024.

[13] Schwarz,H. et al. (1987) Stability studies on derivatives of the bovine pancreatic trypsin inhibitor. Biochemistry, 26, 3544–3551.

[14] Blum, Markus, et al. "InterPro." Version 87.0, EMBL-EBI, 2021, https://www.ebi.ac.uk/interpro/

[15] Pettersen, Eric F., et al. "UCSF Chimera." Version 1.14, University of California, San Francisco, 2004, https://www.cgl.ucsf.edu/chimera/