Alma Mater University of Bologna
Bioinformatics master's degree - 2023
Laboratory of Bioinformatics Project Report

Building a profile Hidden Markov Model to identify Kunitz-type Protease Inhibitor Domains

Nafiseh Barmakhshad

Department of Pharmacy and Biotechnology, FaBiT

Abstract

Kunitz domains are the active domains of proteins that inhibit the function of proteases which are known as biological key regulators of cellular processes. Identification of these domains in protease inhibitors is worthy in many aspects, such as for development of new pharmaceutical drugs.
In this project, a Hidden Markov Model (HMM) is generated and trained by protein sequence datasets and structural alignments to be used for prediction of protease inhibitors by detecting the presence of kunitz domains within a given protein structure.
The model's performance has been tested with already known kunitz-type protease inhibitors and it performed the classification in a sufficient good level of accuracy.

1 Introduction

Kunitz domains are protease inhibitors. They are about 50–60 amino acids long with a molecular weight of about 6 kDa and fold into a disulfide-rich α/β structure. Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI).(1)
The majority of the sequences having Kunitz-BPTI-like inhibitors belong to a family of inhibitors that (3) have one or more inhibitory domains characterized by a conserved spacing between cysteine residues, a typical disulfide binding pattern (2)
This family contains the Kunitz domain which is a common structural fold found in a family of reversible serine protease inhibitors. This domain is thought to have evolved over 500 million years and is ubiquitous in all kingdoms of life and has been incorporated into many different genes. In general, each domain is encoded by a single exon. Some genes encode proteins with a single Kunitz domain, e.g. bovine pancreatic trypsin inhibitor (BPTI), trophoblast Kunitz domain protein (TKDP), amyloid beta-protein precursor (ABPP), as well as Kunitz-type venom peptides such as dendrotoxin. (3)
Elastase-like enzymes are involved in important diseases such as acute pancreatitis, chronic inflammatory lung diseases, and cancer. Structural insights into their interaction with specific inhibitors will contribute to the development of novel anti-elastase compounds that resist rapid oxidation and proteolysis. (4)
Kunitz protease inhibitors are ubiquitous, being found in many organisms, including animals, plants and microbes. According to the MEROPS database, peptidases of the I2 family have been identified in the unicellular choanoflagellate *Monosiga brevicollis* (5).
Plant Kunitz proteins belong to the I3A family of peptidase inhibitors which are unrelated to the Kunitz proteins found in animals (6). At least three families of Kunitz inhibitors have been identified in vertebrates which selectively inhibit serine protease activity and they are involved in various anti-inflammatory processes (7).
In invertebrates, Kunitz proteins, as well as having major rule in serine protease inhibition, some Kunitz proteases can act as ion channel blockers and are known as Kunitz-type toxins (KTT). They are frequent components of the venoms from poisonous animals including snakes, sea anemones (e.g. *Anthopleura elegantissima* (8), cone snails (e.g. *Conus striatus* (9)), tarantulas (*Ornithoctonus* spp.(10)), scorpions (*Lychas mucronatus* (11)) and the cattle tick *Boophilus microplus.* (12)
The Bovine pancreatic trypsin inhibitor (BPTI) is a well-studied model. The fold is stabilized by three characteristic disulphide bridges shown in blue in Figure1.b. Characteristic features of the structures are the presence of six cysteine residues, two anti-parallel beta strands in yellow and a short alpha helix in pink.
profile-HMMs have been extensively used for modelling, protein classification, motif detection due to its convenience and effectiveness in representing sequence profiles.(13) In this study, based on a selection of representative set of protein structures from PDB, a profile Hidden Markov Model was built with HMMER software package. The HMM was tested to assess whether it can be used to find additional homologous in UniProtKB database that may be included within the Kunitz-domain family.

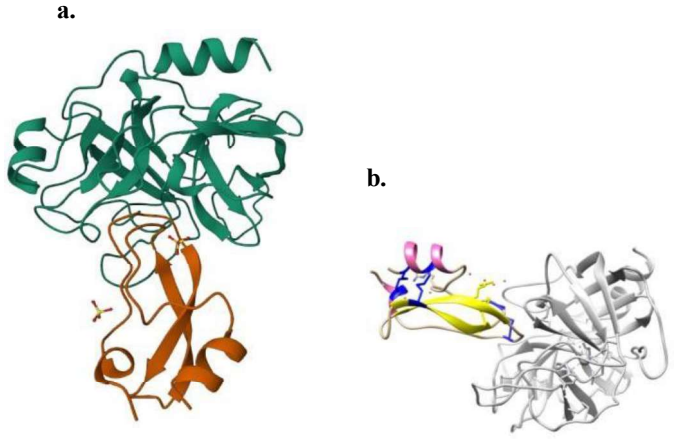

Figure.1 WILD-TYPE RAT ANIONIC TRYPSIN COMPLEXED WITH BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI) from Rattus norvegicus with resolution of 1.80 Å **a)** PDB DOI: https://doi.org/10.2210/pdb3TGI/pdb **b)** The 3 disulfide bonds that keep the structure compact are shown in blue, 2 beta sheets in yellow a little bit rotated but near in the space created by chimera 1.15rc

2 Material and Methods

2.1 Databases

To build the HMM (Hidden Markov Model), a large set of structures were collected from the family of proteins with kunitz domain downloaded from PDB as for this study last update on 05/2023.

To avoid biasing the HMM because of possible database differences, positive and negative control test sequences were performed combining searches on UniProtKB/SwissProt, PDBe and Pfam.

2.2 Structure Selection

Firstly, the RCSB PDB database was queried to fetch a collection of structures with annotated Kunitz domain by Pfam (Pfam AC: PF00014), resolution below 3.0 Å and sequence length between 50 to 80 amino residues inclusive the boundaries. Also, polymer Entities grouped by sequence identity of 95% displayed as Representative. 27 structures were retrieved. I extracted the list containing columns of Entity ID as identifier and PDB ID as Structure Data, and Auth Asym ID as Polymer Entity Data.

2.3 Multiple Structural Alignment

Then after cleaning the list from PDB (Supplementary material-Codes) obtained a clean set of protein structures, the list of PDB codes, to be used in calculation of the multiple structure alignment, given as input to PDBeFold which is an available MSA web tool.

(As reported in the supplementary material-MSA_Results) The Overall RMSD (Root Mean Square Deviation) is 0.7406 which can be considered a fairly good similarity between the input structures. The output file downloaded in Fasta format (Supplementary material-fasta.seq) to be cleaned and prepared for Generation of HMM. In this step the sequences were also checked manually to filter out possible alignment errors such as a shorter sequence than others.

2.4 Generation of the Hidden Markov Model

The Hidden Markov Model was generated by hmmbuild of HMMER (v.3.3.2) on reads of the updated Fasta file obtained previously from PDBeFold. With the Skyalign tool, it's possible to visualize HMMs by making sequence logos as shown in Figure2.

Logos are commonly used in molecular biology to provide a compact graphical representation of the conservation pattern of a set of sequences. They render the information contained in sequence alignments or profile hidden Markov models by drawing a stack of letters for each position, where the height of the stack typically corresponds to the conservation at that position, and the height of each letter within a stack depends on the frequency of that letter at that position.(14)

2.5 Selection of training and testing sets

In order to validate our model, it has been tested with 2 datasets of proteins retrieved by querying SwissProt database for manually annotated sequences with BPTI/Kunitz domain annotated by Pfam (PF00014), so called positive.list (supplementary material), counted as 390 proteins and a set of proteins without annotated Kunitz domain, negative.fasta (supplementary material), counted as 569126 proteins.

To have a fair test, the proteins in the positive set that are the same with the protein structures collected for generating the model presented in the training set, should have been removed. For this goal, since in the training set the PDB-IDs of the proteins were presented but the positive set contains the uniprot-IDs, an ID-Mapping on Uniprot has been performed to extract the corresponding uniprot-IDs so to make the removal of the identical proteins feasible. With comm command the IDs presented only in the positive set filtered in the first column and these IDs were again uploaded in uniprot to the ID mapping section in order to get the Fasta format file of them this time to finally provide it to the HMM and test our model.

2.6 Method optimization and assessment

The hmmsearch program of HMMER was run against all sets with -Z parameter set to 1 and –max option two times, once for the positive set and then for the negative set. In both cases, hmmsearch gave back the sequences for which a Kunitz domain was found with e-values of less than or equal to 1. For the positive set, the result stored in outpot.search file and then a specific portion of this file containing the data of interest cleaned and saved in kunitz_clean.out which has 374 sequences.

For the negative dataset, the same procedure repeated, and the search result file named output_Neg.search has been cleaned in nonkunitz.out file.

(All files are provided in supplementary materials with the corresponding names)

For classifying in further steps, a "0 column" assigned as lacking kunitz domain label, added to the file of sequences without kunitz domain and a "1 column" also added to the file with sequences having kunitz domain respectably named nonkunitz.class and kunitz.class files.

Since nonkunitz.class has only 33 proteins and we need to have the same size as it of the negative file we started with (569126), the e-value of proteins for which the domain was not identified, was manually set to 100 and the missing sequences were added and saved in the recovered_nonkunitz.class.

2.7 Model testing

To test our Model, firstly we needed to know the best threshold of E-value which makes the model performing in it's highest Matthews Correlation Coefficient (MCC), which has been considered as a reliable statistical rate to evaluate binary classifications based on their confusion matrices. A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes. It plots a table of all the predicted and actual values of a classifier.

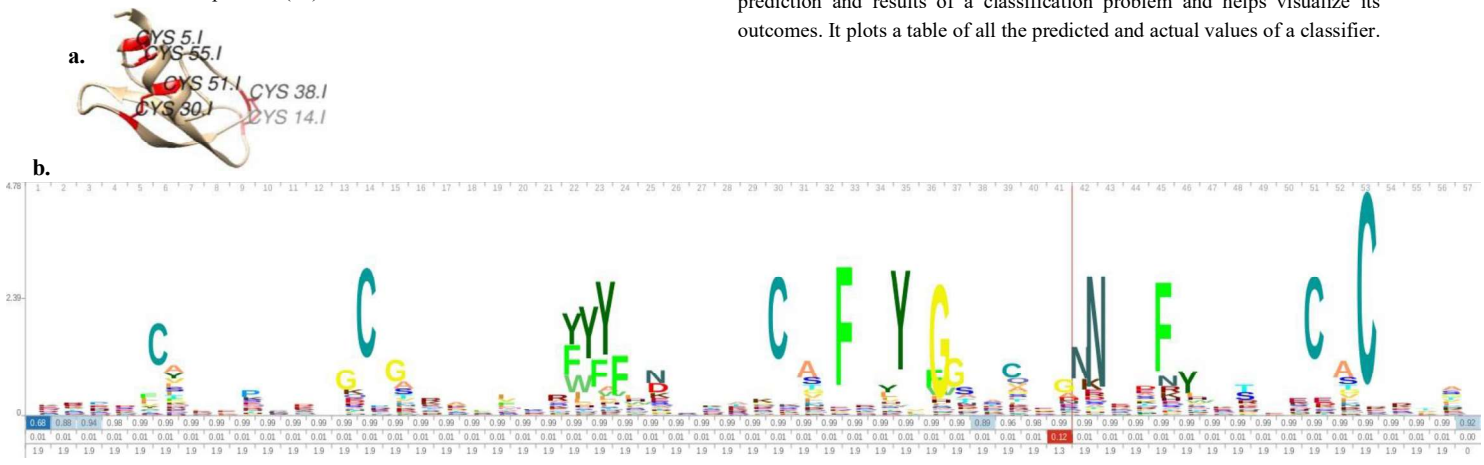

Figure2. *a) **Disulfide Bridges** shown in red between Cysteine 5-55 and 51-30 and 38-14 generated by Chimera1.15rc which can be seen also in b) **Sequence logo of the HMM** created with Skyalign tool and can be compared as in some bridges there are higher information content shown by the bigger size of the Cys residue.*

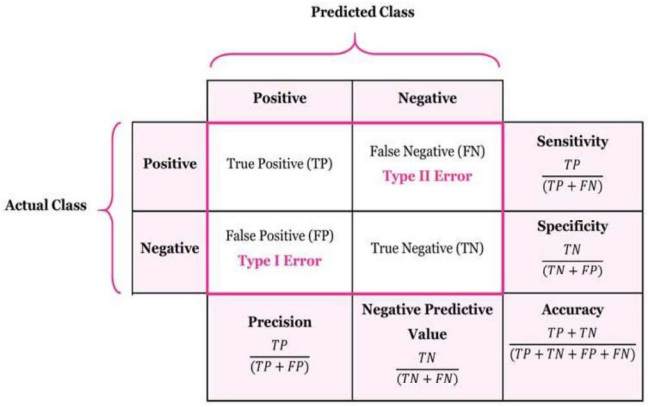

Figure3. *Data Science and Machine Learning_ https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html*

According to figure3, we can define **Sensitivity** as True Positive or Recall and, **Specificity** as Tue Negative Rate which measures the negative examples labeled as negative by classifier. To this end, a 2k-cross validation has been done by creating 2 datasets called training_set.txt and test_set.txt from randomized recovered_nonkunitz.class and kunitz.class obtained in step 2.6 divided by half.

Then the method's performance was evaluated by running the performance.py script (provided in supplementary materials) which returns as the output the **accuracy** (ACC) and the Matthew's correlation coefficient (MCC) for threshold values ranging between 1 and 10^{-20}.

Accuracy is calculated as the total number of two correct predictions of True Positive summed up with True Negative (TP + TN) divided by the total number of a dataset including also all False Positives and False Negatives (TP+ TN + FP + FN) shown in figure.4.a

Matthews correlation coefficient (MCC) is a correlation coefficient calculated using all four values in the confusion matrix as indicated in figure.4.b (15)

a.

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

b.

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

Figure.4.*a) Accuracy Formula – b) Matthews Correlation Coefficient Formula*

After running performance program 3 times, on each training and test sets and a merged of the both files, the highest MCC value has been found by finder.py program.

As shown in table.1 in all the 3 cases the E-value identified as 10^{-8} for which the model guaranteed the best performance with the highest MCC.

	E-value threshold	ACC	MCC
trianing	1e-08	0.9999	0.9945
test	1e-08	0.9999	0.9973
both	1e-08	0.9999	0.9959

Table 1. *The E-values results from the training and test set and both merged files performed corresponding the highest Accuracy and Matthew's correlation coefficient.*

Finally, the hmmsearch has been ran again, setting the E-value threshold to the 10^{-8}, and the result files were cleaned.

As previously known the dataset provided has 569126 nonkunitz in the negative search, the model as expected did not detect any of them over the threshold fixed so it correctly predicted all 569126 of them. But on the other hand, 374 kunits positive domain were assigned in the kunitz_new.search file, but the model detected 370 of these domains over the threshold so it predicted 370 positive. There are 4 input sequences that are not truly positive detected as they should, so our model falsely detected them as negative.

Among 374 positive sequences our model detected 370 of them correctly and did not assign any real negative sequence to positive ones so 0 false positive.

	Positives	Negatives
Predicted Positives	TP = 370	FN = 4
Predicted Negatives	FP = 0	TN = 569126

Table 2. *Confusion matrix of kuniz domain detection model with E-value set to 10^{-8}*

The 4 missing uniprot-ids as False Negatives were detected as: D3GGZ8, O62247, P86963 and Q11101.In the Discussion part each of them are analyzed by details to know the probable reasons that the model did not detect them.

3 Results

3.1 ROC plot visualization

The receiver operating characteristic (ROC) curve, is the plot of the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting.

$$\text{True Positive Rate (TPR)} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

also called sensitivity/recall/hit rate

$$\text{False Positive Rate (FPR)} = \frac{FP}{N} = \frac{FP}{FP + TN}$$

also called fall out

The ROC Area Under the Curve (AUC) score tells us how efficient the model is. The AUC value is within the range [0.5–1.0], where the minimum value represents the performance of a random classifier and the maximum value would correspond to a perfect classifier.(17)

Here as we can calculate the TRP is 0.9893 and FPR is 0.

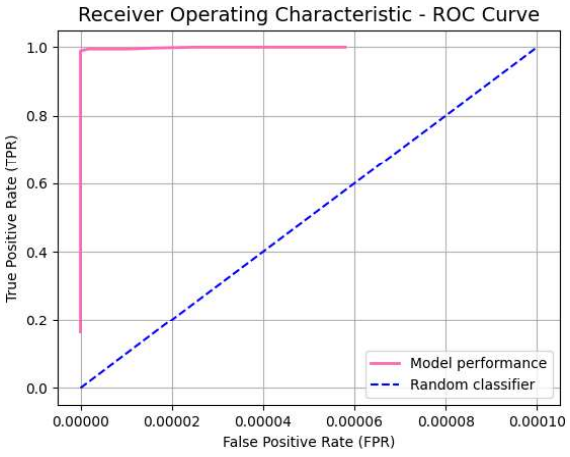

Figure.5 *The comparison of model receiver operating characteristic(ROC) curve with a random classifier.*

As shown in Figure.5, the ROC curve of the model used in this study shows a classifier with an almost perfect performance level.

4 Discussion and Conclusion

To conclude, as the aim of this project was to build a Hidden-Markov Model to predict presence of kunitz domains of any given protein sequences, the model has been trained with known datasets and tested with both having and lacking the kunitz domain protein datasets from SwissProt and evaluated.

Although the evaluation measurements indicate a perfect model performance, still there are some false negatives which have been reported as kunitz domain container protein sequences from reviewed uniprot, but our model did not detect them. Following, each of the four proteins were checked.

D3GGZ8, Kunitz-type protein bli-5 from Haemonchus contortus (Barber pole worm) with annotation score of 2/5 is a protein inferred by homology which means that it's existence is probable because orthologs exist in closely related species and not proven experimental evidences.

Besides, there is a curator inference evidence in a publication, claiming this protein appears to have serine protease activity in vitro (PubMed:19716386). However, it is uncertain if this activity is genuine as bli-5 lacks all the catalytic features of serine proteases. So maybe the model was not so irrelevant in not detecting this protein as positively Kunitz-type.

For **O62247**, Kunitz-type protein bli-5 from Caenorhabditis elegans, although the annotation score is as high as 5 and the protein existence been proved by evidence at protein level, there are 2 publications with manual assertion based on experiment indicating that they appear to lack serine protease inhibitor activity in vitro when tested with bovine pancreatic alpha-chymotrypsin and elastase. (PubMed:16500660, PubMed:19716386)

Also given that there are no similar proteins at 90% identity for this isoform, maybe for this case also we can say the model is sensitive on the serine protease inhibitor activity and lack of it in the sequence of the protein leaded to be not detected as positive.

Finally, **P86963**, named BPTI/Kunitz domain-containing protein-2 from Pinctada maxima (Silver-lipped pearl oyster- White-lipped pearl oyster) and also **Q11101**, BPTI/Kunitz inhibitor domain-containing protein-C02F12.5 from Caenorhabditis elegans, both have annotation scores of 2/5 with protein existence based on evidence at protein level and none of them has similar proteins at 90% identity for this isoform. But more importantly, checking PROSITE-ProRule Annotation of each of the P86963 and Q11101 for their kunitz domains, the function is indicated as "Undefined", so this is being in consistent with what was hypothesized as for the other 2 proteins and lead us to assume the model is more sensitive to the function of kunitz domain rather than its only its existence.

Since for this project, the model has been tested firstly on March 2023 with 386 positive proteins and the uniprot website was updated on the following months, so later at the end of May, the final version of the model tested with the positive filter set as from Pfam- PF00014 and reviewed, contained 390 this time and it changed the number of false positive results, probably the 4 false positives of the model are the new added proteins.

About Multiple Structural Alignment described in part 2.3, it is suggested for more analysis instead of only using the PDBeFold website to download the MSA file, one may try to obtain it also by running MTM-align program as a free available tool and compare the results.

Also following this study, some more features can be added such as detecting the location of kunitz domain in terms of the amino acids positions in the protein structure so it can be more precisely used, but this may need more vast majority of data to train the model.

5 References

1. Kunitz domain. In: Wikipedia [Internet]. 2021 [cited 2023 May 14]. Available from: https://en.wikipedia.org/w/index.php?title=Kunitz_domain&oldid=1042869443
2. Ikeo K, Takahashi K, Gojobori T. Evolutionary origin of a Kunitz-type trypsin inhibitor domain inserted in the amyloid ? precursor protein of Alzheimer's disease. J Mol Evol. 1992 Jun
3. CDD Conserved Protein Domain Family: Kunitz-type [Internet]. [cited 2023 May 14]. Available from: https://www.ncbi.nlm.nih.gov/Structure/cdd/cl00101
4. García-Fernández R, Perbandt M, Rehders D, Ziegelmüller P, Piganeau N, Hahn U, et al. Three-dimensional Structure of a Kunitz-type Inhibitor in Complex with an Elastase-like Enzyme *. J Biol Chem. 2015 May 29;290(22):14154–65.
5. Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem. 2004
6. Enzymically Catalyzed Disulfide Interchange in Randomly Cross-linked Soybean Trypsin Inhibitor | Elsevier Enhanced Reader [Internet]. [cited 2023 May 14]. Available from: https://reader.elsevier.com/reader/sd/pii/S0021925818970031?token=31E5CF5A017EB71148393740FCBB899F18F0C8C4C073741CA18B134D03AFC061CFC0D2D366C1FFCEC81A1516B3633C62&originRegion=eu-west-1&originCreation=20230514155759
7. Shigetomi H, Onogi A, Kajiwara H, Yoshida S, Furukawa N, Haruta S, et al. Anti-inflammatory actions of serine protease inhibitors containing the Kunitz domain. Inflamm Res. 2010 Sep
8. Peigneur S, Billen B, Derua R, Waelkens E, Debaveye S, Béress L, et al. A bifunctional sea anemone peptide with Kunitz type protease and potassium channel inhibiting properties. Biochem Pharmacol. 2011 Jul 1;82(1):81–90.
9. Dy CY, Buczek P, Imperial JS, Bulaj G, Horvath MP. Structure of conkunitzin-S1, a neurotoxin and Kunitz-fold disulfide variant from cone snail. Acta Crystallogr Sect D. 2006;62(9):980–90.
10. Yuan CH, He QY, Peng K, Diao JB, Jiang LP, Tang X, et al. Discovery of a Distinct Superfamily of Kunitz-Type Toxin (KTT) from Tarantulas. PLOS ONE. 2008 Oct 15;3(10):e3414.
11. Zhao R, Dai H, Qiu S, Li T, He Y, Ma Y, et al. SdPI, The First Functionally Characterized Kunitz-Type Trypsin Inhibitor from Scorpion Venom. PLOS ONE. 2011 Nov 8;6(11):e27548.
12. Lima CA, Torquato RJS, Sasaki SD, Justo GZ, Tanaka AS. Biochemical characterization of a Kunitz type inhibitor similar to dendrotoxins produced by Rhipicephalus (Boophilus) microplus (Acari: Ixodidae) hemocytes. Vet Parasitol. 2010 Feb 10;167(2):279–87.
13. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep;10(6):402–15.
14. 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. Basic evaluation measures from the confusion matrix [Internet]. Classifier evaluation with imbalanced datasets. 2015 [cited 2023 May 29]. Available from: https://classeval.wordpress.com/introduction/basic-evaluation-measures/
16. Park SH, Goo JM, Jo CH. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. Korean J Radiol. 2004;5(1):11–8.
17. Bhandari A. Guide to AUC ROC Curve in Machine Learning : What Is Specificity? [Internet]. Analytics Vidhya. 2020 [cited 2023 May 30]. Available from: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Supplementary Material

All the codes and files used in this project are provided in the following GitHub link:

https://github.com/nafisabr/LB1-final-project.git

Since the negative.fasta file was too big to be uploaded in GitHub (569126), it can be downloaded here:

https://www.uniprot.org/uniprotkb?query=NOT%20(xref:pfam-pf00014)%20AND%20(reviewed:true)