Building HMM-Based model for kunitz domain prediction and evaluation

Behrouz Mollashahi[1]

[1]Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, 40126 Bologna, Italy.

Abstract

Introduction: Kunitz domain is a serin protease domain exerting various functions within the cells. They are widely studied in plants, animals, and humans. Various studies show their importance in various fields. The aim of this study is to build a Hidden Markov Model (HMM) that can
predict the Kunitz domain and evaluate the model.

Method: In this study, data was extracted from Protein Data Bank (PDB) website under constraints and then Multiple structural Alignment (MSA) was performed with the PDB-fold website, the results were then cleaned and used to generate HMM model with the HMMER package. Finally, the evaluation of the model was done with the accuracy test and Matthew Correlation Coefficient (MCC).

Result: Based on testing the generated HMM model with the Swiss-prot database the high MCC score shows a significant performance of the HMM model generated for the Kunitz domain.

1. Introduction

A coagulation inhibitor called tissue factor pathway inhibitor (TFPI) controls the onset of coagulation brought on by tissue factor. TFPI is an essential regulatory element in regulating the effects of tissue factor in experimental models (1). Mice lacking the TFPI gene experienced intrauterine coagulopathy and vascular disintegration, which caused fetal death. The TFPI immune depletion in rabbits significantly lowered the tissue factor coagulation threshold. On the other hand, high-dose exogenous recombinant TFPI may raise this threshold and prevent venous thrombosis and disseminated intravascular coagulation(2). TFPI contains two Kunitz domain which is the factor Xa is inhibited by the Kunitz-2 domain of TFPI, while the Kunitz-1 domain inhibits the factor VIIa/tissue factor catalytic complex in a factor Xa-dependent manne(3). (figure 1).

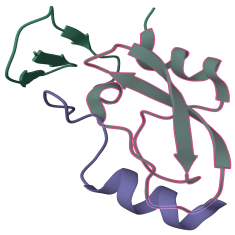


Figure1: The kunitz domain

Serine proteases and their inhibitors are molecules of interest when pursuing therapeutic applications since they are involved in a wide range of cellular processes (for example, coagulation, inflammation). The family of serine protease inhibitors featuring a Kunitz type motif is one that is frequently explored. Due to their biological plasticity and biotechnological usefulness, they are widely distributed in nature and the focus of research in a variety of domains (such as plants, human health, and parasitology)(4). Numerous studies link proteins with Kunitz domains to immune system evasion by various helminth parasites, despite their inhibitory effect against trypsin, chymotrypsin, elastase, and other serine proteases(5,6).

In this work, we attempt to create a statistical model for protein with Kunitz domain discrimination. We did this by beginning with members of this protein family who already had experimental structural data accessible. Following that, we created the relevant sequence profile based on a Hidden Markov Model (HMM) to serve as the model's seed utilizing multiple structure alignment. Finally, we analyzed our model, and the findings showed that HMM performed admirably.

2. Methods

2.1 Data wrangling and preparation:

We extract the highly annotated proteins with accessible experimental structure in order to build the seed alignment from the PDB database. The PDB IDs and their associated information on the Kunitz domain were obtained from PDB databases. Accordingly, the data of the Kunitz family (PF00014), which included PDB IDs, PDB chains extracted based on the following filters: Data collection resolution was set to less than 3 and, the Pfam family identifier was set to PF00014 to retrieve only proteins containing the Kunitz domain. Since each domain should have about 50 residues, we also include a 49-100 length filter on the residues length to exclude any non kunitz parts of the protein. For the cleaning and processing the downloaded data from PDB, we used linux terminal. (Supplementry 1).

2.2 Multiple Structural Alignment:

Multiple Structural Alignment exerted on the cleaned data using PDB-Fold web server (https://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver).
The identifiers of the selected Kunitz domains downloaded under the Auth Asym ID from PDB website advanced search, the file included the identifiers then cleaned and prepared as input for the pdb-fold (supplementary 1). On PDB-fold the parameter set to multiple and the cleaned identifiers uploaded to build MSA matrix. To evaluate the results of the PDB-Fold, multiple sequence alignment was performed on the output of the PDB-fold.

2.3 Model Generation:

Based on the MSA results, the aligned sequences are retrieved and The aligned Kunitz domain sequences are used to generate a profile HMM using software package HMMER v3.3.2 (supplementry 1). A HMM profile is a probabilistic model that represents the variability of the aligned kunitz domain sequences at each position in the sequence and this probability is visualized with Skyalign website

4. Model Evaluation:

The profile HMM is tested against a set of positive and negative Kunitz domain sequences to evaluate its sensitivity and specificity. The sensitivity measures the ability of the model to correctly identify true kunitz domain sequences, while the specificity measures the ability of the model to reject non-kunitz domain sequences. In order to do that we created two datasets:

1. A negative set made up of all UniProtKB/Swiss-Prot entries that aren't Kunitz entries and contains 566628 sequences.
2. A group of positive entries from UniProtKB/Swiss-Prot that all include the Kunitz domain. Since we used some of entries in this dataset used for model construction, we built a query excluding those that have PDB structure associated to them, resulting in 336 sequences.
3. We assigned zero to the negative and one to the positive sets and then combined the two sets together. Then datasets were shuffled and divided into half and used the cross-validation resampling approach. Therefore, we can train and analyze the model with different E-value criteria on separate iterations on each sets to get the greatest accuracy (ACC) and Matthew correlation coefficient (MCC). The average threshold of each set was used to calculate the final, comprehensive threshold.

1 Results

3.1 Data wrangling and preparation:

PDB searches yielded 190 sequences, respectively. However, only 33 entries were selected for the seed construction after the curation and filtering process.
The PDB-Fold algorithm produced a satisfactory alignment of the sequences based on the result's matrix showing RMSD scores which are very low(<1), that indicate the perfect alignment of the Kunitz domain. Also the results of the multiple sequence alignment on the sequences downloaded as the output of the pdb-fold indicated the residues have good conservation, especially, the cysteine disulfide bridges (figure 2). Therefore it can be inferred the PDB-fold Multiple structural alignment were satisfactory.

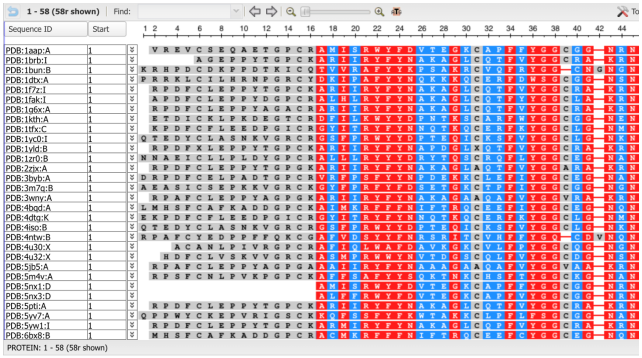


Figure.2: Multiple sequence alignment of kunitz domain by NCBI sequence alignment viewer

3.2 HMM Model Generation

Hmmbuild was able to create a model that could describe the position-based emission and transmission probability given the profile of multiple sequence alignment. Figure 3 displays the model's logo. By creating a model based on average match state emission probability.

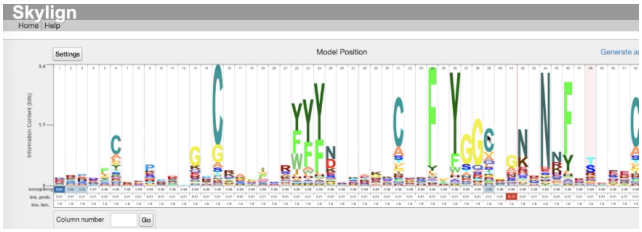


Figure.3: Model's logo of the kunitz domain

3.4 Model Evaluation

We evaluated the accuracy and Matthew correlation coefficient of our model by establishing a reviewed negative and positive set with 566629 and 336 elements, respectively. The test exerted on the first set gives the best threshold of 0.001 after each iteration(Table.1a). The same test exert on set two and the results were shown in Table.1b. Then the average value based on the two thresholds calculated (0.01) and the best performance MCC considered accordingly (MCC:0.9906, ACC:0.9999). Based on the average threshold the model shows 7 false negatives in the classification (table 2). To assess the performance, Receiver Operating Characteristics curve is plotted(figure.4) which also shows Area Under the Curve (AUC) > 0.5 in terms of specificity and sensitivity of our model, meaning that the performance of the model has better than random classification.

Set 1 Training — Set 2 Performance evaluation.

TH	ACC	MCC		TH	ACC	MCC
0.1	0.9999718901749134 MCC	0.978931242				
0.01	0.9999859450874566 MCC	0.989297549				
0.001	0.9999894588155925 MCC	0.991940946		0.0001	0.9999859451368417 MCC	0.989196628
0.0001	0.9999894588155925 MCC	0.991940946				
1E-05	0.9999859450874566 MCC	0.989196628				
1E-06	0.9999859450874566 MCC	0.989196628				
1E-07	0.9999824313593209 MCC	0.986444702				
1E-08	0.9999824313593209 MCC	0.986444702				
1E-09	0.9999824313593209 MCC	0.986444702				
1E-10	0.9999824313593209 MCC	0.986444702				

Table.1a: set 1 used as training and set 2 used to evaluate the performance

Set 2 Training — Set 1 performance evaluation.

TH	ACC	MCC		TH	ACC	MCC
0.1	0.9999859451368417 MCC	0.989297549				
0.01	0.9999894588526312 MCC	0.991940946		0.01	0.9999859450874566 MCC	0.989297549
0.001	0.9999859451368417 MCC	0.989196628				
0.0001	0.9999859451368417 MCC	0.989196628				
1E-05	0.9999859451368417 MCC	0.989196628				
1E-06	0.9999859451368417 MCC	0.989196628				
1E-07	0.999982431421052 MCC	0.986444702				
1E-08	0.999982431421052 MCC	0.986444702				
1E-09	0.999982431421052 MCC	0.986444702				
1E-10	0.999982431421052 MCC	0.986444702				

Table.1b: set 2 used for training and set 1 used to evaluate the performance

E-Value	Sequence	Description
4.60E-24	sp|C0HLA7|MAMB4_DENVI	Mambaquaretin-4 OS=Dendroaspis vir
5.00E-24	sp|C0HLA5|MAMB2_DENV	Mambaquaretin-2 OS=Dendroaspis vir
1.10E-23	sp|C0HLA6|MAMB3_DENPO	Mambaquaretin-3 OS=Dendroaspis pol
2.30E-23	sp|C0HLA8|MAMB5_DENJA	Mambaquaretin-5 OS=Dendroaspis jam
1.10E-22	sp|C0HLA9|MAMB6_DENVI	Mambaquaretin-6 OS=Dendroaspis vir
2.40E-14	sp|C0HLB0|MAMB9_DENV	Mambaquaretin-9 OS=Dendroaspis vir
0.0026	sp|P84555|TIQ7_RHISA	Kunitz-type serine protease inhibi

Table2: Showing the false negative proteins. The ones that has Kunitz domain but classified wrongly as a nonkunitz.

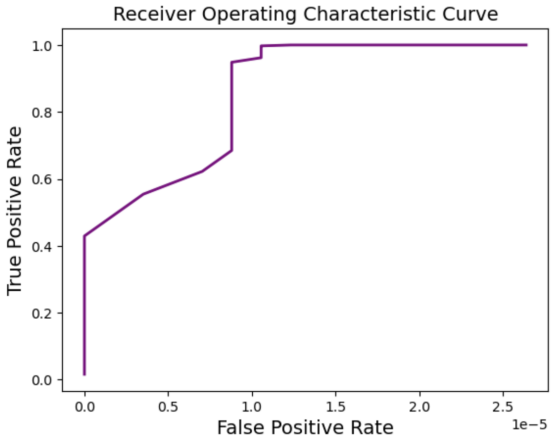
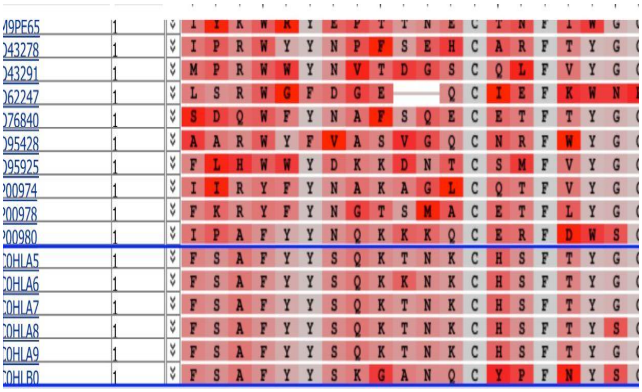


Figure.4: ROC curve for evaluating the specificity and sensitivity of the model.

To assess further the false negative results, since 6 of the false negatives belongs to the venom proteins in western

snack (dendroaspis polylepis polylepis), the sequences of all this 6 proteins for the kunitz domains aligned with other kunitz domain proteins (figure 5), the results suggest that there is a level of variability between these 6 venomes' proteins and the others in their kunitz domain.



Figure(5): showing that the venome's snack Kunitz protein are more close to each other rather than other Kunitz.

4. Discussion:

Kunitz domain is a serine-protease protein that is well studied. Their role in many diverse cellular functions such as host cellular inflammatory response, coagulation are well approved. Kunitz domain has length of 50 to 60 amino acids and 6kD weight that can be folded into its disulfide rich alpha/beta structure. Hidden Markov Model is a statistical model in which we are able to model protein domains based on alignment of sequence or structure of protein families. In this project we use HMM for modeling the Kunitz domain based on structural alignment and check for the reliability of the method.

It is safe to believe that the model can identify proteins with some degree of precision based on the existence of the domain, given the findings that were achieved. The positive dataset might be more accurate if more cross-references for the domain were taken into account, hence minimizing the need to deal with the manual search of the misclassifications, which might be more time-consuming when working with bigger datasets. In total we can choose for our model's the threshold of 0.01 with accuracy of 0.999, MCC of 0.990, with 7 false negatives. 6 of the false negetives belongs to the snake's venome. it can be suggested since snake venomes has biologically rapid evolution and diversification(7,8), it can result in levels of sequence variation that might not be well represented in the training set to be captured by HMM model.

References

1. Mast AE, Ruf W. Regulation of coagulation by tissue factor pathway inhibitor: Implications for hemophilia therapy. J Thromb Haemost JTH. 2022 Jun;20(6):1290–300.
2. Huang ZF, Higuchi D, Lasky N, Broze GJ. Tissue factor pathway inhibitor gene disruption produces intrauterine lethality in mice. Blood. 1997 Aug 1;90(3):944–51.
3. Dockal M, Hartmann R, Fries M, Thomassen MCLGD, Heinzmann A, Ehrlich H, et al. Small Peptides Blocking Inhibition of Factor Xa and Tissue Factor-Factor VIIa by Tissue Factor Pathway Inhibitor (TFPI) *. J Biol Chem. 2014 Jan 1;289(3):1732–41.
4. Patel S. A critical review on serine protease: Key immune manipulator and pathology mediator. Allergol Immunopathol (Madr). 2017;45(6):579–91.
5. Kowalczewska ME, Brożyna A, Jóźwicki W, Pławski K, Przybyszewski M, Wrotek S, et al. Analysis of the involvement of cytokines in allergy and breast cancer association. Contemp Oncol Poznan Pol. 2014;18(6):396–402.
6. Yang SN, Hsieh CC, Kuo HF, Lee MS, Huang MY, Kuo CH, et al. The effects of environmental toxins on allergic inflammation. Allergy Asthma Immunol Res. 2014 Nov;6(6):478–84.
7. Accelerated evolution of crotalinae snake venom gland serine proteases - ScienceDirect [Internet]. [cited 2023 May 31]. Available from: https://www.sciencedirect.com/science/article/pii/S0014579396011441
8. Serrano SMT. The long road of research on snake venom serine proteinases. Toxicon. 2013 Feb 1;62:19–26.