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

Ahmad Alkhansa

May 6, 2019

Abstract

Kunitz inhibitors are divided into two protien families one that belong to the plantae kingdom (PF00197)(1) and the other to animalae kingdom (PF00014)(1). The focus of this article is on the Bovine pancreatic trypsin inhibitor domain (BPTI) (pfam ID:PF00014). This protein has a clinical significance in blood coagulation that can lead to reduced probability in blood transfusion during surgeries. The approach to annotate the domain is by building a profile hidden markov model (HMM) which will predict the existence of such domain in newly discovered genes. To let this succeed, I collected the best domain presenters by entering protein database and acquiring the highest quality structures. To decrease the redundancy of the same structure, clustering method is adopted. Those structures are the seeds for the Hidden Markov Model. I tested the model by setting the domain expected value threshold deduced from a training set. The training set and the testing sequences are obtained and processed from the Uniprot Swissprot Database. The results of the predictor show high success in separating the protiens that contain the Kunitz domain from the those that lack it. The performance of this predictor is of a very high quality.

1 Introduction

Bovine pancreatic trypsin inhibitor domain (BPTI) (pfam ID:PF00014)(2) is a functional domain that represents the kunitz-type protease inhibitors in the animal kingdom. They are small protiens (**Fig. 1**) with average weight of 6 KDa and a relative length between 50 and 60 amino acidic residues(1). The structure consists of a small hydrophobic core, antiparallel beta sheets, alpha helices and 3 disulfide bridges that stabilizes the structure. These Inhibitors are involved in nitric oxide synthase type-I and -II action inhibition and the prevention K + transport by Ca 2+ -activated K + channels. In addition, they react with serine proteases to form stable structures. The clinical importance of the BPTI is that it reduces the production of plasmin, lowers the production of hemorrhagic complications and the need of blood transfusion(2).

2 Materials And Methods

2.1 Datasets

The resources that were essential for data collection: Protien Data Bank(3) (PDB) and Uniprot Swissprot(4). PDB contains

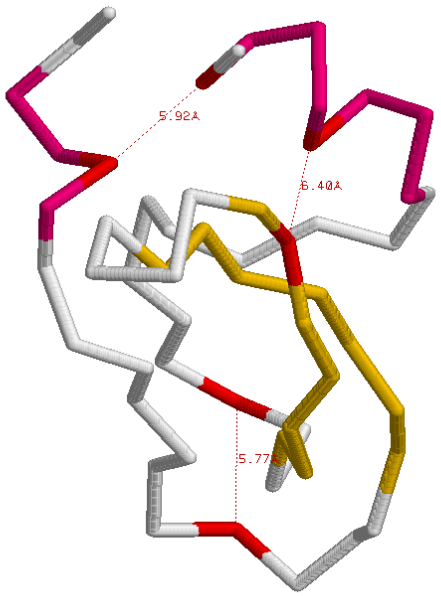

Fig. 1: A graphical representation using rasmol(8) of a BPTI functional unit (PDB ID 1B0C). The helices are marked in Purple and the sheets are marked in Yellow. The cysteines that form the disulfide bridges are colored Red.

the necessary sequences to build the Hidden Markov Model. There are two results from the PDB database. The first is through the advanced search which is a table of 161 rows with the following columns: ID, chain length, experimental method, resolution, chain ID and entity ID. The second result is through PDBefold software. The software undergo pairwise structural alignment between a query of my selection which is 5PTI and the whole PDB database. The file contains the alignment values (Z-scores, Q-scores …) and the 161 PDB IDs with their chains. I created a new file containing the merged IDS and the chains with the resolution. 19 seed structures were used to build the model after clustering and removing redundancy. Uniprot Swissprot provided the positive and the negative sets needed for training and testing the model. The database provided 500 kunitz negative sequences for training and 547897 for testing the model in addition to the 343 positive sequences.

2.2 Procedure and Methods

2.2.1 Building Profile HMM

The first step in building the Hidden Markov Model is by selecting its seeds. I set the parameters of PDB advanced search by selecting structures of pfam(5) ID PF00014, maximum resolution of 2.0 angstrom, without artificial mutation, X-ray diffraction as experimental method and since a typical kunitz protien is of length between 50 and 60 amino-acidic residues, the preferred length is between 50 and 70. The PDB result showed several protiens with different resolutions. The focus is on the domain with the highest resolution so the structure with this character is of PDB ID 5PTI. I used PDBeFold online software for structural alignment to undergo a pairwise alignment between 5PTI and the whole PDB database to collect as much similar structures as possible. Blastclust used the structures' sequences in both PDB search and PDBeFold results to cluster those entities. The clustering is necessary to reduce redundancy hence the bias of the highest number of most similar structures during the building of the HMM. This is done by taking a structure with the highest resolution from each of the 19 clusters. Those are considered the seeds of HMM. The Second step of building the HMM is by preparing a multiple structural alignment (MSA)(**Fig. 6**). PDBeFold is used again for the MSA of the seeds to prepare the

file for the hmmbuild software of HMMER(7) package. The output file contains the profile of the kunitz domain (**Fig. 2**).

2.2.2 HMMsearch

The profile HMM let me distinguish proteins that contain Kunitz domain from the others. This can be done by running the hmmsearch software of HMMER against several sets. In the first time, I prepared 2 FASTA files obtained from the Uniprot Swissprot database, the first one is Kunitz positive and the other is Kunitz negative. The procedure of the positive file preparation is by applying the advanced search and setting the following parameters: sequences that belong to the kunitz family (PF00014) and are reviewed. In the result, I selected to download the tabular format that contain the Uniprot entries and their corresponding PDB IDs. To eliminate the sequences that belong to the seeds of the HMM, I removed the seed sequences from the file using a shell command. I downloaded the Swissprot sequence
Database and extracted the sequences of the positive file using a python script. I ran hmmsearch and used the HMM and the positive file as inputs. The output is a tabular file that contain the domain's E-value that I am interested in. In the preparation of the negative file, the selection was upon protiens that does not contain the kunitz domain with at least 45 amino-acidic residues (kunitz type protiens have length between 50 and 60).
I selected randomly 500 ids and extracted the sequences. This time, the only difference in the hmmsearch input is the negative file instead of the positive one but the output has the same format. To make the domain E-values comparable, I normalized by dividing them by the related set size. In addition each result is labeled to prepare them for the confusion matrix. The same procedure is done for the rest 547897 negative sequences.

2.2.3 Performance Measurement

Labeling the results and merging the files is essential for the formation of the confusion matrix. The python script of the confusion matrix distinguish the true from false kunitz domain positive and negative by reading the labels and the E-values in the merged file in addition to an input which is E-value threshold. The threshold is estimated by the observation the E-values inside the positive and the negative files. In fact, E-value threshold is considered as a necessary reference for domain recognition.
Performance evaluation of the model relies on three crucial calculations whichare the true positive rate (TPR), the false positive rate (FPR) and the matthews correlation coefficient (MCC) (**Table. 1**).

E-value	TPR	FPR	MCC
1.00E-01	1.0	3.60E-02	0.1283046448
1.00E-02	1.0	0.0033236174	0.3974619102
1.00E-03	1.0	0.0002920256	0.825656959
1.00E-04	1.0	4.75E-05	0.9641033061
1.00E-05	1.0	9.13E-06	0.9927855706
1.00E-06	0.9970845481	5.48E-06	0.9941866022
1.00E-07	0.9941690962	5.48E-06	0.9927184697
1.00E-08	0.9941690962	3.65E-06	0.9941654459
1.00E-09	0.9941690962	3.65E-06	0.9941654459
1.00E-10	0.9912536443	3.65E-06	0.9926972262

Table. 1 The different e-value thresholds provide different True positive rate, False positive rate and matthews correlation coefficient.

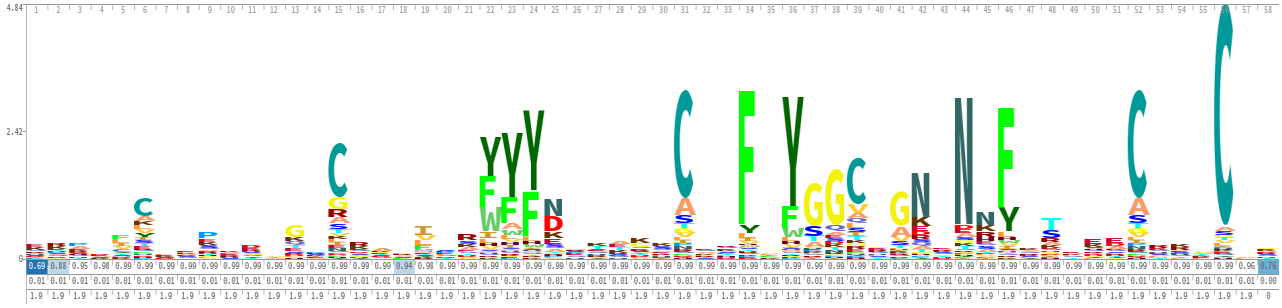

Fig. 2: The Logo of the profile HMM using skylign(6). The 6 Cysteines that are necessary for the kunitz-type protien structure exist significantly in this profile.

TPR (**Fig. 3**) is the ratio of the positives detected (true positives in the confusion matrix) by the model from the total positives. The model may falsely recognize negatives as positives (false positive in the confusion matrix), the rate of such event happening in the negative set is expressed as FPR (**Fig. 4**). MCC (**Fig. 5**) is a measurement of agreement between the model and the observation, it ranges between -1 and +1. The highest value means total agreement while -1 means the contrary. However, 0 values resemble a random predictor.

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Fig. 3: The formula that calculates the true positive rate (TPR). It's the ratio of true positive (TP) over total positives (P) which is the sum of the True positive (TP) and False Negative (FN).

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

Fig. 4: The division of false positive (FP) by the total negative (N) that consist of false positive (FP) and true negative (TN).

$$MCC = \frac{TP*TN - FP*FN}{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]^{(1/2)}}$$

Fig. 5: Matthews Correlation Coefficient (MCC). The formula includes the true positives and negatives, in addition to the false positives and negatives

Fig. 6: Multiple sequence alignment of the seed sequences visualization using Seaview(2).

3 Results and Discussion

Multiple possibilities for the E-value threshold showed different results. In **Table. 1**, the True positive rate is equal to 1 between E-value threshold of 1e-01 and 1e-05 . However, it starts decreasing as the threshold gets smaller. The false positive rate is continuously decreasing until it reaches 1e-08 where the rate stops changing. The optimum E-value threshold is 1e-05 since it keeps the True positive rate at its highest value with relatively very low False positive rate and the MCC is in its highest value at this point.

Adopting the optimum E-value threshold gave no false negative but 5 false positive out of 547892 negatives (**Table. 2**).

The five false positive sequences are of the following Uniprot IDs: C0HLB2, G3LH89, P56409, P40500 and P78746. C0HLB2 and G3LH89 are Kunitz-type serine protease inhibitors and belong to BPTI-like superfamilies which explain the high similarity.

4 Conclusion

In Conclusion, the HMM is an accurate profile that can be used as a predictor and annotation tool for the Kunitz domain.

True Postive	343	False Negative	0
False Positive	5	True Negative	547892

TPR	1.0	E-value threshold	1.00E-05
FPR	9.13E-06	MCC	0.99278557062
Precision	0.929539295393		

Table. 2 Confusion Matrix result showing True/False positives and negatives, True positive rate (TPR), False positive rate (FPR), Precision, E-value threshold and Matthews correlation coefficient (MCC).

Reference

(1) '*Analysis of Kunitz inhibitors from plants for comprehensive structural and functional insights'* International Journal of Biological Macromolecules 113 (2018) 933–943 DOI: 10.1016/j.ijbiomac.2018.02.148

(2) '*The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein'* Current Protein and Peptide Science, 2003, 4, 231-251 DOI : 10.2174/1389203033487180.

(3) '*The Protein Data Ban*k' H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) *Nucleic Acids Research*, 28: 235-242. doi:10.1093/nar/28.1.235

(4) '*The universal protein resource (UniProt)*' Nucleic Acids Research, Volume 28, Issue 1, 1 January 2000, Pages 235–242 DOI: 10.1093/nar/gkm895

(5) '*The Pfam protein families database in 2019'* Nucleic Acids Research (2019), DOI: 10.1093/nar/gky995

(6) '*Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models*' Published: 13 January 2014, DOI: https://doi.org/10.1186/1471-2105-15-7

(7) '*HMMER web server: interactive sequence similarity searching*' Published: 2011 May 18. DOI: 10.1093/nar/gkr367

(8) '*RasMol: Biomolecular graphics for all'*, *Trends in Biochemical Sciences (TIBS)*, September 1995, Vol. 20, No. 9, p. 374. DOI: https://doi.org/10.1016/S0968-0004(00)89080-5

(9) '*SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.'* 2010 Feb;27(2):221-4. doi: 10.1093/molbev/msp259.

Supplementary material

The workflow exists in a repository with scripts, sequences and a pdf workflow file. The link is:
https://github.com/ahmadalkhansa/bioinformatics_masters_ahmad_alkhansa/tree/master/second_semester/LB1_second/projects/report