Profile Hidden Markov Model for the Kunitz-type domain

Biagio Iacolare
University of Bologna, Master degree in bioinformatics

Abstract

Motivation: Kunitz domains are the active domains of proteins that are able to recognise specific protein structures and are involved in the inhibition of the function of protein degrading enzymes (proteases). Kunitz-type proteins are involved in different biological processes, such as the reduction of bleeding during surgery. After this remark came the intention to develop molecules including this domain as pharmaceutical drugs. The aim of the project is to automatically annotate Kunitz domain in Swiss-Prot building a Hidden Markov Model based on structural alignment. **Results**: The profile HMM was built using multiple tools and software and than statistical validated using different sets of data. Model prediction has been optimized, obtaining a model that was able to identify effectively Kunitz-type domains in the dataset, showing an accuracy of approximately 100% and a Matthews Correlation Coefficient equal or very close to 1.

Introduction

Kunitz domains are the active domains of proteins that inhibit the function of protein degrading enzymes or, more specifically, domains of Kunitz-type are protease inhibitors. They are relatively small, with a length that range between 50 to 60 residues and a molecular weight of 6 kDa. Because of its activity the Kunitz-domain has been exploited to develop important biopharmaceutical drugs. An important example is Aprotinin a serine protease inhibitor, also known as Bovine Pancreatic Trypsin Inhibitor (BPTI). It is an antifibrinolytic molecule that inhibits trypsin and related proteolytic enzymes slowing down fibrinolysis and thus reducing bleeding. The physiological function of BPTI determines the protective inhibition of the major digestive enzyme trypsin when small amounts of it are produced by the cleavage of the trypsinogen precursor during storing in the pancreas. The structure of BPTI-like domains is a disulphide rich alpha+beta fold [1] (Fig. 1) characterised by the conservation of Lys/Arg 15, important for the function of BPTI-like inhibitors, and Cys residues (at positions 5, 14, 30, 38, 51, 55) involved in three disulphide bridges, that are important for structure stabilization [2]. A profile Hidden Markov Model for the Kunitz-type domain was generated using available structural information in order to annotate Kunitz domain in Swiss-Prot. The dataset of protein sequences has been compared to the profile HMM and, based on the statistical significance of the alignments, protein sequences are assigned or not to the Kunitz-type family, obtaining the confusion matrix of the dataset analysed. Information included in the confusion matrix are then used to determine important parameters to assess the reliability of the model as accuracy and Matthews correlation coefficient.

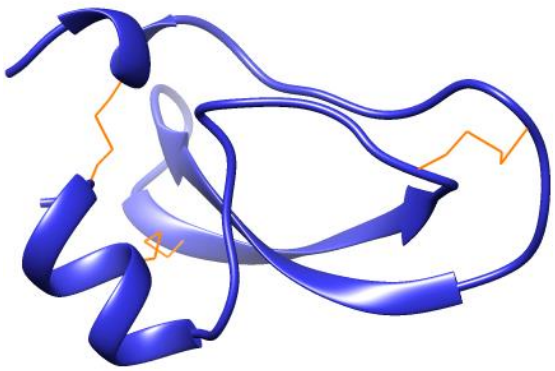

Fig. 1
In blue 3D structure of Kunitz/Bovine pancreatic trypsin inhibitor domain. In yellow lateral side chains of Cysteines and disulphide bridges.

Materials and Methods

HMM model generation

In order to build Hidden Markov Model we downloaded from RCSB PDB [3] a *csv file* with sequences of protein containing Kunitz domain (PFAM identifier PF00014). We only downloaded sequences of structures characterised by high resolution (lower than 3.5 Å) and by a number of polymer residues lower than 100. We modified the *csv file* to obtain a *fasta file* containing only

sequences with length between 50 and 70 residues. Then, to remove redundancy, we used *blastclust*, a program within the BLAST package [4], to cluster protein sequences. The program uses the *blastp algorithm* to compute pairwise matches. We specified 0.99 both as score threshold and coverage threshold to cluster together sequences with 99% of identity and 99% of coverage. The program returned a file containing 14 clusters. After a file containing only the first element of each cluster was created, we performed the structural alignment of the selected domains using PDBeFold multiple 3D alignment service [5] (results_PDBefold.txt, Supplementary) providing as input the list of non-redundant PDB codes. In the output *fasta file* containing the multiple sequence alignment we deleted 2FJZ due to its high RMSD value (3.0360 Å) and low Q-score (0.1842). The HMM model is generated using HMMER [6] *hmmbuild command*. It reads a multiple sequence alignment file to build a new Profile HMM, that will be saved in a hmmfile. The Profile HMM is a probabilistic model that includes information about evolutionary changes that have occurred in a set of related sequences. Then, we used Skylign [7] to generate the HMM logo of the model (Fig. 2) that provides the graphical representation of the conservation pattern of profile HMM. In an HMM logo, under the sequence profile, three rows are added to indicate the frequencies of occupancy (presence) and insertion, as well as the expected insertion length.

Validation of HMM prediction

In order to validate the HMM prediction, we retrieved from UniProtKB [8] a positive dataset of manually curated proteins containing the BPTI/Kunitz domain and a negative dataset of manually curated proteins not containing the BPTI/Kunitz domain. The positive set is composed by 359 proteins. To remove redundancy, we run a blast search using blastpgp [9]. Then from the positive set we removed the 5 sequences that showed 100% of identity with the sequences used to build the model, obtaining a positive dataset with 354 sequences. The negative set is composed by 561,894 proteins, which do not contain Kunitz domain. To perform the two-fold cross validation of the model, both datasets were divided into two halves, a training set to train the model and a testing set to test the trained model. We validated the dataset against the model using *hmmsearch* [6], that reads the Profile HMM from hmmfile and searches for significantly similar sequence matches in the dataset. The output of the *hmmsearch* is a file containing sequence hits with statistical significance based on E-value. We created, both for training and testing set, two files containing 3 columns reporting the following elements: UniProt identifier, associated E-value and, in the first file, number 1 to indicate proteins that belong to the positive set, whether in the second file number 0 to indicate proteins belonging to the negative set. Finally, using the *comm* command we added in the final file, that contains the UniProt identifiers of negative sequences retrieved by *hmmsearch*, all the sequences excluded by the tool because of the too high E-value. Then we assigned them an E-value equal to 10 (for the final datasets used in this project check the folder Datasets in Supplementary). To measure the performance of our model we used a program (performance.py, Supplementary), that takes as input the merged file of positive and negative datasets containing the entries with the associated E-value and number 0 or 1 to identify the dataset of belonging. The program returns confusion matrices based on different E-value thresholds and the related accuracy (Acc) and Matthews correlation coefficient (MCC). Accuracy represents the proportion of correct prediction, while Matthews correlation coefficient is used in machine learning to measure the quality of binary classifications. These two parameters are thus used to measure the goodness of our model in discriminate proteins belonging to the Kunitz-type family.

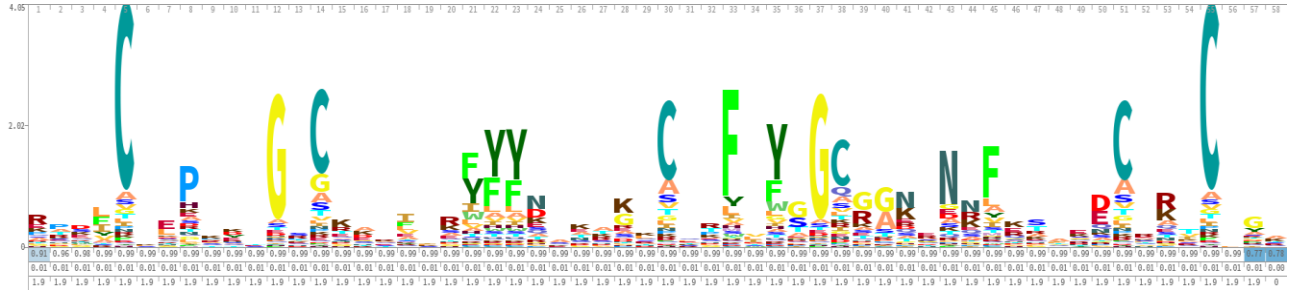

Fig. 2
HMM logo of the model obtained using hmmbuild on sequence alignment file obtained from PDBeFold. It is possible to observe conservation of the most important residues in Kunitz domain, in particular the six cysteines involved in disulphide bridges important for structure stabilization.

Results and Discussion

The model is first applied on the training set. Between the different analysed thresholds, we selected 1e-07 as optimal threshold because of the maximized accuracy and Matthews correlation coefficient (Table 1). With this threshold we obtained a perfect classification for the training set, lacking in false negatives and false positives (Table 2). After having set 1e-07 as threshold for the discrimination between positive and negative proteins, we applied the model on the testing set obtaining again good results, but with a protein recognised as false positive and two proteins recognised as false negatives (Table 3).

	Threshold	Accuracy	MCC
Training	1E-06	1,0	1,0
	1E-07	1,0	1,0
	1E-08	0,99999643082	0,99716935936
	1E-11	0,99999286165	0,99433068084
Testing	1E-06	0,99998936437	0,99150007877
	1E-07	0,99998936437	0,99150007877
	1E-08	0,99998936437	0,99150007877
	1E-11	0,99997872875	0,98291077393

Table 1. Performance of the Profile HMM with different E-value thresholds applied on training and testing set. As result of the analysis 1e-07 in the training set has been selected as optimal threshold showing accuracy 1,0 and MCC 1,0.

	Kunitz	Non-Kunitz
Predicted Positive	TP=177	FP=0
Predicted Negative	FN=0	TN=280000

Table 2. Confusion matrix computed on the training set using as E-value threshold 1e-07.

	Kunitz	Non-Kunitz
Predicted Positive	TP=175	FP=1
Predicted Negative	FN=2	TN=281893

Table 3. Confusion matrix computed on the testing set using as E-value threshold 1e-07. In this case the prediction is not perfect showing 1 false positive and 2 false negatives.

Analysing results coming from the confusion matrix, computed on the testing set, we identified the UniProt entries D3GGZ8 and O62247 as false negatives. They belong to the Kunitz-type family, but our model was not able to recognise them as positives. These two proteins derive respectively from *Barber pole worm* and from *Caenorhabditis elegans*, thus, because of the evolutionary distance, they show a low sequence similarity with the Profile HMM, that was generated using predominantly proteins derived from Bos Taurus. On the other hand, G3LH89 was retrieved as false positive. Further analysing G3LH89 in UniProtKB, we found that it belongs to Bombus ignitus and possess a BPTI/Kunitz inhibitor domain despite the lack of the Pfam code correspondent to it. For this reason, our model was able to correctly assign this protein to predicted positive proteins belonging to Kunitz-type family.

Conclusions

In conclusion, exploiting the HMMER package, it was possible to obtain a Profile HMM based on structural information to automatically annotate Kunitz domain in Swiss-Prot. The model generated has proven to be enough reliable in recognising such domain in the database, with high level of accuracy and good Matthews correlation coefficient.

References

[1] https://en.wikipedia.org/wiki/Aprotinin
[2] Current Protein and Peptide Science, 2003, 4, 2 The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein ≠ P Ascenzi et. al
[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne.
(2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.
[4] http://ftp.ncbi.nih.gov/blast/documents/blastclust.html
[5] E. Krissinel and K. Henrick (2004). *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions*
[6] HMMER v3.3: http://hmmer.org/
[7] Wheeler, T.J., Clements, J., Finn, R.D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics Volume 15 (2014) p.7 DOI: 10.1186/1471-2105-15-7
[8] The UniProt Consortium UniProt: a worldwide hub of protein knowledge Nucleic Acids Res. 47: D506-515 (2019)
[9] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410