Hidden Markov Model build for the detection of the Kunitz domain

Paolo Mastrogiovanni[1]

[1] Master in Bioinformatics, University of Bologna, Bologna
E-mail: paolo.mastrogiovann2@studio.unibo.it

Abstract

Motivation: The Kunitz/BPTI family of proteins, with their capability to inhibit the serin proteases, have demonstrated to be important in many biological processes, such as angiogenesis and coagulation. To have tools that can characterize the domain it's needed for the further understanding of its role, and its involvement in other interactions. HMMs are probabilistic models extensively used in the biological field, especially for the prediction of protein sequences. Because of their reliability in predicting consecutive residues, they were chosen as tool for the analysis of the Kunitz domain. All the process that led to the build of the model can be found here. More specific information are in the supplementary material, with the full detailed pipeline.

Results: The Hidden Markov Model that was built in this study has proved to be able to correctly classify proteins based on the presence of the kunitz domain. The model, with a 99% of MCC and Accuracy, has produced just a few errors, that indeed were also seen to be borderline cases. For this reason, it's safe to say that the model can be used for the study of the domain.

Supplementary materials: https://github.com/pmastrogiovanni/kunitz_hmm

Keywords: HMM, Kunitz, protein-domains

1. Introduction

Peptides of the Kunitz/BPTI family contain one of the most evolutionarily ancient and conserved structural motifs, the Kunitz fold, which is widely distributed among both terrestrial and marine organisms. Historically, the firstly discovered representative of this family, the bovine pancreatic trypsin inhibitor (BPTI), is known as an inhibitor of different serine proteases and capable of carrying out an anti-inflammatory function participating in proliferation and angiogenesis(Gladkikh *et al.*, 2020). In fact, Kunitz/BPTI homologs are, based on their three-dimensional (3D) structure, classified into two families: canonical Kunitz-type inhibitors including BPTI-like toxins, and anticoagulant proteins. The latter have secondary structures whose orientation is clearly similar to that of BPTI, but differ in the folds of some loops and particularly in the orientation of the N-terminal segment. Most of the canonical Kunitz-type homologs inhibit serine proteinases of the chymotrypsin family through their highly conserved anti-proteinase site (Župunski, Kordiš and Gubenšek, 2003). The domain comprises around 60 amino acid residues and is characterized by three highly conserved disulphide bridges with the bonding patterns 1-6, 2-4, and 3-5(Dai, Zhang and Huang, 2012).

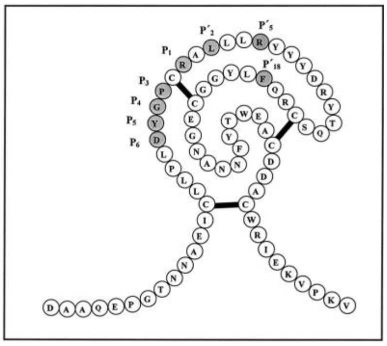


Fig. 1 - Model structure of the first Kunitz-type domain of human TFPI-2

It consists of a protruding loop (L1, the reactive site) that, in protease inhibitors, binds to the catalytic-site cleft of cognate proteases, predominantly via the side chain of the central residue (named P1). The L1 loop is usually preceded by a short 310 helix at the N-terminus and is always followed by a twisted antiparallel β-hairpin. To complete the fold, a short, C-terminal α-helix (α1) is connected to the β-hairpin by a second loop (L2) and positioned next to the 310 helix at the ”base” of the structure (Paesen *et al.*, 2009)(Fig.1). The standard mechanism of inhibition involves a strong and non-covalent interaction, like the complex that would be formed between the enzyme and its substrate. The Kunitz-BPTI inhibitors directly block the active site of the targeted serine proteases. The loop formed by the domain, with his convex and prolonged structure, is exposed to the solvent and is highly complementary to the enzyme concave active site. Residues that precede or follow this segment or from more remote regions can also participate in the interaction and influence the energy of association (Bomediano Camillo *et al.*, 2021).

2. Methods

1.1 Datasets curation

1.1.1 Training set

The training set was obtained starting both from a PDB (Berman *et al.*, 2000, May-2022) and a PDBeFold (Krissinel and Henrick, 2004) research. To find proteins containing the kunitz domain on the PDB’s advanced search tool, the pfam identifier PF00014 (id for the BPTI/Kunitz domain) was used as filter. Additional filters were added to clean the results, avoiding structures with more than 3Å as resolution and proteins with polymer entity mutations. Upon the research, 125 hits were found plus 7 which were identified as artifacts, hence were excluded. PDBefold is an online tool that can be used to perform both pairwise or multiple structural alignment. To find proteins with the kunitz domain, a pairwise alignment search was performed, using the 3tgi protein and it’s “I” chain (domain bearing) as reference structure. The tool was run with 70% as match threshold, and produced 652 results, which were filtered taking only the proteins with a Z-score ≥ 3 and a RMSD ≤ 1.5Å, resulting in 336 IDs. The two protein datasets that were obtained by these researches were merged, and the correspondent sequences were taken from the pdb_seqres fasta file. In order to reduce redundancy, the cd-hit software (Fu *et al.*, 2012) was used with a cut-off of 0.95, to clusterize all sequences with an identity of 96% or more in groups, selecting a representative sequence for each. The alignment that was used to build the hmm was produced by the PDBeFold multiple structural alignment tool on the set of 21 clusters.

1.1.2 Cross-validation set

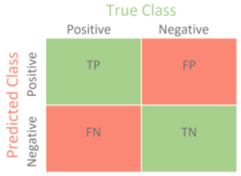
For the optimization of the model, the cross-validation set was obtained from the reviewed Uniprot KB (Bateman *et al.*, 2021). The 336 positives were produced using the pfam identifier PF00014 as filter, taking all the proteins which were not cross-referenced to the PDB, to obtain a completely different dataset from the training one. The 557267 negatives were obtained upon the search for proteins without the pfam identifier reference, and characterized by a sequence length between 40 and 10000 to reduce computation time. Each set was randomly sorted and split in two subsets, which were then used to test the model and find the optimal e-value.

2.2 HMM build

The Hidden Markov Model was generated using the software package HMMER(v3.2.2), which takes in input a multiple alignment to build the profile for the domain (command hmmbuild). The HMM can then be used to compute the probabilities of a given set of sequences to be generated by the model. Sequences bearing the Kunitz domain are reported to have a low e-value and are classified as positives. To compute the optimal e-value for the threshold, the model was tested on the cross-validation subsets using the command hmmsearch (see github repository for the specific parameters). The HMM Logo was produced using the Skylign open software (Wheeler, Clements and Finn, 2014).

2.3 Evaluation metrics

A customized python script (accuracy.py, on the github repository) was used to compute the performances of the model. The script takes in input a file storing a set of proteins with the associated e-values and class membership, and given a threshold, computes the confusion matrix, the accuracy, and the Matthews Correlation Coefficient (MCC)(Fig.2).



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

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

Fig. 2 - In order: Confusion Matrix, Accuracy and MCC formulas

Because of the imbalance between the classes of positives and negatives, the MCC was chosen as reference performance metric since the accuracy tends to be more sensible to the imbalances. Therefore, the goal of the optimization phase was to find the optimal e-value threshold to maximize the MCC. An e-value range between 1e-3 and 1e-10 was tested on each of the cross-validation subset, to take then the average between the ones resulting in the highest MCC. The threshold was used for the labelling of the sequences: proteins with an e-value below the threshold were labelled as positives, while the ones above it were considered negatives. The comparison between the predicted labels and the actual class memberships can be found in the confusion matrix (Fig. 4), storing True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

3. Results and discussion

To build an efficient Hidden Markov Model it's important to have a very well curated dataset, with high reliability on the sequence of the target domain. For this reason, the training set was obtained through cross-research with both the PDB and the PDBeFold, filtering out to take only the proteins with high quality resolved structures and devoid of mutations. An additional step could be to search also directly on the Pfam database, to obtain all the sequences that are reported to have the kunitz domain. The training dataset was clusterized, and then aligned, such that it could be used by the HMMER software to make the HMM. The HMM Logo, representative of the domain, can be found in the figure 5, where it's possible to recognize the elevated presence of cysteines, needed for the disulphide bonds. To optimize the model, the sets of positives and negatives were downloaded from the Uniprot KB, to be then elaborated, producing two subsets, which were needed for the 2-fold cross-validation. A range of e-values between 1e-03 and 1e-10 was tested on both subsets, and, after averaging the thresholds that maximized the Matthews Correlation Coefficient, the selection of the optimal one was made (Fig.3).

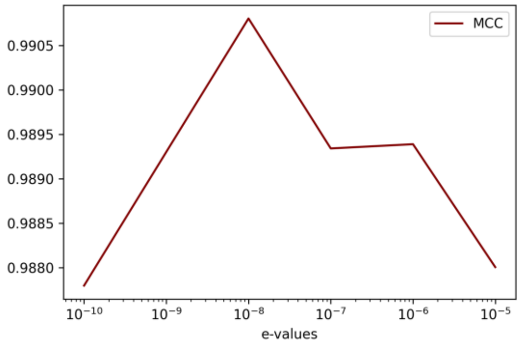


Fig. 3 - Plot of the MCCs obtained with multiple e-value treshold on the full cross-validation dataset. Made with Matplotlib. E-values below 1e-5 excluded for clarity.

The selected threshold was 1e-08 since it gave a MCC of 0.9908. The confusion matrix reports 4 false positives and 2 false negatives (Fig.4), which ID were retrieved to analyse manually their Uniprot entries. The 4 negatives that were labelled as positives by the model (FP), are all isoforms of the PI-stichotoxin protein, which is a serin protease. On the Uniprot entries for these cases (P0DV05, P0DV04, P0DV03, P0DV06), it's possible to see how the proteins are referred to as Kunitz bearing (Sintsova *et al.*, 2015). Looking at the cross-references in fact, a manually annotated prosite identifier for the domain can be found, but none associated to pfam. Since the pfam ID was used as discriminant for the Uniprot search, the 4 proteins were considered part of the negative dataset. Based on this result, it's reasonable to say that, possibly, they are not actually false positives, and the classification asserted by the model it's correct. Searching then the false negatives on the Uniprot database (O62247, D3GGZ8), it's evident that they are borderline cases. Even if they actually have a pfam identifier related to the kunitz domain, as stated from the associated paper, they appear to have the function of inhibiting the serine proteases, but it is uncertain if this activity is genuine, as the protein lacks all the catalytic features of the serine proteases (Stepek, McCormack and Page, 2010). Even if it is not possible to know for certain the actual class of the two proteins, the misclassification made by the model it's understandable.

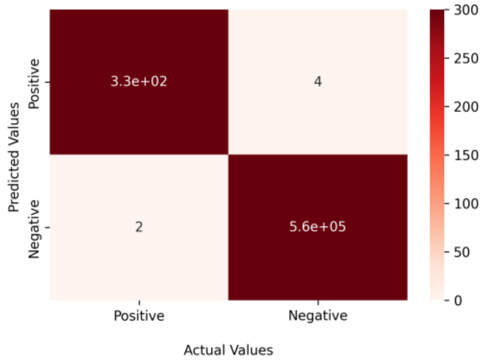


Fig. 4 - Representation of the confusion matrix in a Seaborn Heatmap Plot. Colour range goes from 0 to 300 for graphic purposes.

4. Conclusions

Based on the results that were obtained, it's safe to assume that the model is able to classify the proteins based on the presence of the domain, with a certain precision. Possibly, by taking into consideration other cross-references for the domain, there could be more precision in the positive dataset, avoiding the need to deal with the manual search of the misclassifications, which could be more time expensive when working with larger datasets. However, the Hidden Markov Models are confirmed to be dependable for this type of tasks and remains one of the most efficient computational methods for protein domain analysis.

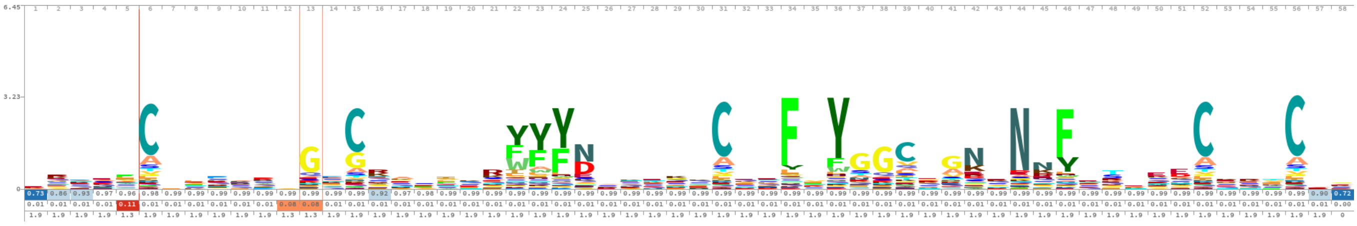


Fig. 5 - HMM Logo generated with Skylign

References

Bateman, A. *et al.* (2021) ‘UniProt: the universal protein knowledgebase in 2021’, *Nucleic Acids Research*, 49(D1), pp. D480–D489.

Berman, H. M. *et al.* (2000) ‘The Protein Data Bank’, *Nucleic Acids Research*, 28(1), pp. 235–242.

Bomediano Camillo, L. de M. *et al.* (2021) ‘Structural modelling and thermostability of a serine protease inhibitor belonging to the Kunitz-BPTI family from the Rhipicephalus microplus tick’, *Biochimie*, 181, pp. 226–233.

Dai, S. X., Zhang, A. Di and Huang, J. F. (2012) ‘Evolution, expansion and expression of the Kunitz/BPTI gene family associated with long-term blood feeding in Ixodes Scapularis’, *BMC Evolutionary Biology*, 12(1), p. 4.

Fu, L. *et al.* (2012) ‘CD-HIT: accelerated for clustering the next-generation sequencing data’, *Bioinformatics*, 28(23), pp. 3150–3152.

Gladkikh, I. *et al.* (2020) ‘Kunitz-Type Peptides from the Sea Anemone Heteractis crispa Demonstrate Potassium Channel Blocking and Anti-Inflammatory Activities’, *Biomedicines*, 8(11), pp. 1–17.

Krissinel, E. and Henrick, K. (2004) ‘Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions’, *urn:issn:0907-4449*, 60(12), pp. 2256–2268.

Paesen, G. C. *et al.* (2009) ‘An Ion-channel Modulator from the Saliva of the Brown Ear Tick has a Highly Modified Kunitz/BPTI Structure’, *Journal of Molecular Biology*, 389(4), pp. 734–747.

Sintsova, O. V. *et al.* (2015) ‘Anti-inflammatory activity of a polypeptide from the Heteractis crispa sea anemone’, *Russian Journal of Bioorganic Chemistry*, 41(6), pp. 590–596.

Stepek, G., McCormack, G. and Page, A. P. (2010) ‘The kunitz domain protein BLI-5 plays a functionally conserved role in cuticle formation in a diverse range of nematodes’, *Molecular and Biochemical Parasitology*, 169(1), pp. 1–11.

Wheeler, T. J., Clements, J. and Finn, R. D. (2014) ‘Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models’, *BMC Bioinformatics*, 15(1), pp. 1–9.

Župunski, V., Kordiš, D. and Gubenšek, F. (2003) ‘Adaptive evolution in the snake venom Kunitz/BPTI protein family’, *FEBS Letters*, 547(1–3), pp. 131–136.