Building a profile Hidden Markov Model for the recognition of Kunitz-type domain in proteins

Davide Abbondandolo

University of Bologna, Msc in Bioinformatics

Abstract

Motivation: Kunitz-type proteins are a family of protease inhibitors. These proteins have found great pharmaceutical interest thanks to their ability to reduce bleeding in surgery. In this work we built and validate an HMM for the recognition of this domain.
Results: The obtained HMM can predict quite well the Kunitz-type domain in proteins. The measure performed on training and testing sets shows an accuracy near 100% and an MCC value close to 1. These results indicate high reliability and the few mistakes are mostly due to problems in the SwissProt annotation.
Contact: davide.abbondandolo2@studio.unibo.it
Supplementary information: Dataset and script are available at this link.

1 Introduction

Kunitz-type proteins are an important group of ubiquitous protease inhibitors. These proteins can have single or multiple Kunitz domains linked together or associated with other domain types[1]. Their Pfam database[2] (v 32.0) family, Kunitz_BPTI (PF00014), includes 15055 sequence and 275 structures. One of the most iconic members of this class is Aprotinin (bovine pancreatic trypsin inhibitor, BPTI). It is a monomeric globular peptide of 58 residues that folds into a stable, compact tertiary structure with 3 disulfides bonds, a twisted β-hairpin and a C-terminal α-helix. The 3 disulfide bonds are fundamental for the protein stability and the Cysteines involved (Cys5-Cys55, Cys14-Cys38, Cys30-Cys51) are conserved across the whole family (Figure 1). The mechanism of inhibition involves the exposed side chain of Lysine 15 which uses a tight, non-covalent interaction to block the serine protease active site without any conformational changes. Aprotinin is used as a drug and commercialized by Nordic Group under the name Trasylol[3]. Due to its ability of slowing down fibrinolysis, it is administered by injection for prophylactic use to reduce bleeding during complex surgery, leading to a decreasing need of blood transfusions. Profile Hidden Markov Models are one of the most useful tools for modelling, motif detection and classification in proteins[4]. In this work we use a set of Kunitz domain structures to build its profile-HMM. Moreover, we used annotated sequences to validate it.

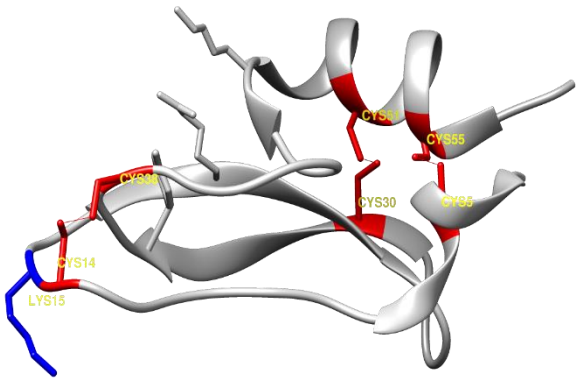

Figure 1: 1BPI structure, where Cysteines involved in disulfides bond are highlighted in red and the inhibition Lysine in blue.

2 Material and methods

Databases

We used RCSB PDB[5] (Release: 2020_04) database to retrieve a representative set of protein structure with the Kunitz domain to build the HMM

and UniProtKB[6] (Release: 2020_02) to download sequences of Kunitz-type and not Kunitz-type proteins to check the model performance.

Validation dataset

The protein sequence on which we will test the model are retrieved from UniProtKB. The reviewed proteins with the Pfam annotation for the Kunitz domain constitute the positive set, which is formed by 359 sequence. The negative set is composed by the remaining 561894 reviewed sequence.

HMM generation

To retrieve the sequence with the Kunitz domain we perform an advanced search on RCSB PDB with the following parameters:

- Resolution <= 3.5 Å
- Pfam annotation = “PF00014"
- Number of Polymer Residues < 100

From this research we obtain 39 structures. We download their sequences and since we know that the length of the domain is 58 residues, we further filter those results to keep only sequences between 50 and 70 residues. This should allow us to keep only sequences that represent the domain itself. On those sequences we perform a clustering procedure with Blastclust, program included in the Blast Legacy[7] (v 2.2.26). The program uses the blastp algorithm to compute pairwise alignment and put the protein in the same cluster if they reach a given threshold of sequence identity and of length coverage given by the user. We set them to 95% and 90% respectively. With those values we can keep into account the sequence domain variability without risking overrepresentation. This procedure returns 8 clusters and from them we select the best element in term of resolution (Table 1) and use PDBefold[8] (v.2.58) to do a multiple structural alignment.

Cluster member	Structure selected for the alignment
1BPI:A 1BPT:A 1BTI:A 1FAN:A 1NAG:A 4PTI:A 5PTI:A 6PTI:A 7PTI:A 8PTI:A 9PTI:A	1BPI:A
2FJZ:A 2FK1:A 2FK2:A 2FMA:A	-
1G6X:A 1K6U:A 1QLQ:A	1G6X:A
1KNT:A 1KTH:A 2KNT:A	1KTH:A
6Q61:A	6Q61:A
30FW:A	30FW:A
5YV7:A	5YV7:A
1DTX:A	1DTX:A

Table 1: Cluster formed by Blastclust and structures seleted for the structural alignment

The alignment evaluation matrices return good values of RMSD, Q-score and sequence identity for all the possible pair, expect for the one involving a member of the 2° cluster (2FJZ:A, 2FK1:A, 2FK2:A, 2FMA:A), so we remove it. Another analysis without it give better alignment results and so we use it to build the HMM. The model is generated by hmmbuild, a program part of HMMER[9] (v.3.3). It reads the multiple sequence alignment file and generates a profile of it in a .hmm file. Its graphical visualization can be obtained with the online tool Skylign[10], that uses the .hmm file to produce a sequence logo (Figure 2). It is noteworthy that the Cysteines involved in the first and third disulphide bridge (Cys5-Cys55, Cys30-Cys51) have higher information content than the second one.

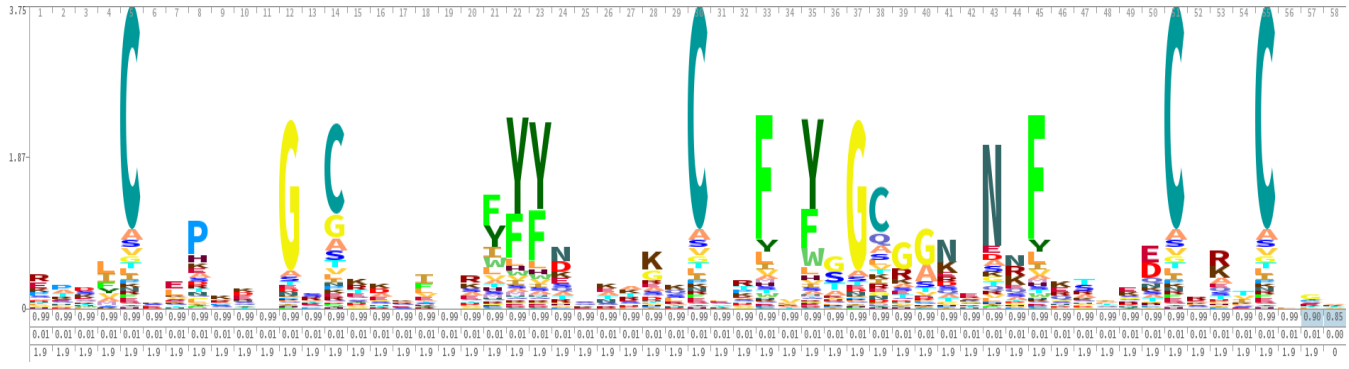

Figure 2: Sequence logo of the Kunitz domain based on our model

Model validation

Once we got our model, we want to evaluate its ability to distinguish proteins with a Kunitz-domain – the positive set – from proteins which do not present it – the negative set. Before doing it, to avoid bias is important to filter out from the positive set the sequences that are too similar to the ones that have been used to build the model. To do so, we make a database out of the sequence used to construct the model with makeblastdb, then we use blastpgp to look in it for hit with a low e-value and 100% of sequence identity in our positive set. The 5 hits are removed from the positive set, then we proceed with a 2-k fold cross validation. We randomize the order of sequences in positive and negative sets, then split them in two halves. hmmsearch is used to align the sequences in the sets with the profile. We use parameters –max to remove all the heuristic filter and -z 1 to normalize the e-value results. Every hit has a score and e-value <= 10 for the whole sequence and the best domain. Sequences that do not reach this threshold are missing, so we manually add them with an e-value of 100. Positive and negative sets are then divided, and their halves are assigned to set1 and set2.

Measure of performance

Performances are automatically computed with a script, that for e-value threshold in the range from 1 to 10^{-20} returns the confusion matrix for the set and calculate the accuracy (AC), which is the ratio of correct prediction to total predictions made, and Matthews correlation coefficient (MCC), which measure the quality of binary classifications (Figure 3). The best threshold is the one that returns the higher value using set1 as training and set2 as testing and is confirmed by switching the set role.

Accuracy (AC)

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

Matthews correlation coefficient (MCC)

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

Figure 3: Formulas to compute accuracy and Matthews correlation coefficient

3 Result and discussion

The developed model performs at its best with an e-value threshold between 10^{-8} and 10^{-10} while using both set1 or set2 as training. The obtained confusion matrices are shown in Table 2.

Set 1

	Positives	Negatives
Predicted Positives	177	1
Predicted Negatives	0	280947

Set 2

	Positives	Negatives
Predicted Positives	175	0
Predicted Negatives	2	280947

Table 2: Confusion matrix with the best threshold for the two set.

Considering set1 as training, the accuracy is higher than 0.999 and the Matthews correlation coefficient is higher than 0.997. The results are also good for the testing set, in which we have an AC higher than 0.999 and the MCC higher than 0.994 (Table 3).

	Set1		Set2	
E-value	AC	MCC	AC	MCC
10^{-8} – 10^{-10}	0.999	0.997	0.999	0.994

Table 3: AC and MCC value for the two set

Those results are really satisfying and allows us to say that our model succeed in predicting whether a protein contain a Kunitz-type domain. Further validation is obtained computing the ROC curve (Figure 3), a performance measurement in which the true positive rate (TPR) is plotted against the false positive rate (FPR).

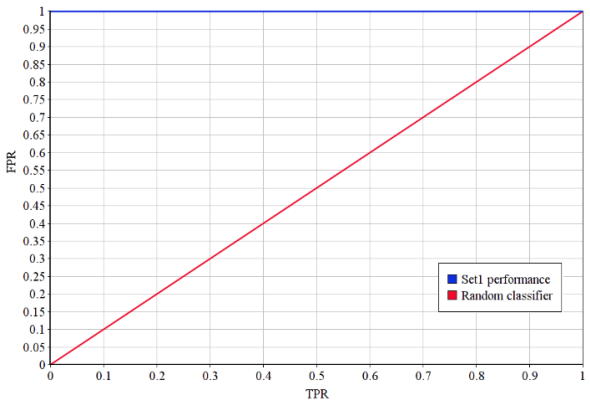

Figure 4: ROC curve of the training Set compared with the expected result of a random classifier.

Although the good results, it is important to point out the presence of 1 False Positive from set1 (UniProtKB AC: G3LH89) and 2 False Negatives from set2 (UniProtKB AC: D3GGZ8, UniProtKB AC: O62247). Those outliers are detected for peculiar reasons which are worth to be discussed.

Kunitz-type serine protease inhibitor Bi-KTI (G3LH89) is a protein with a fully functional Kunitz domain and this can be also highlighted with the alignment with the model, that shows the conservation of all its key residues. However, the Pfam label for the domain, PF00014, is missing from the UniProt entry and this leads to the classification into the false positive. This is one of the main problems of our workflow, which heavily relies on the UniProt annotation. Mistake in it can have a huge impact on the overall quality of the results. D3GGZ8 and O62247 are protein with a Kunitz domain but they achieve low e-value ($1.9x10^{-5}$ and $1.6x10^{-6}$ respectively). This is probably given by the inability of hmmsearch to find the Kunitz domain in those sequence, due to a low sequence identity. If we look at the alignment between their best scoring domain and the consensus sequence of the HMM it is easy to spot that many key residues are not conserved (Figure 5).

4 Conclusion

Our results prove the model ability to predict the presence of a Kunitz-type domain in proteins. Although some limitations given by the protein annotations and some struggle to detect the correct starting point of the domain with a low sequence identity, we can overall be satisfied with it. Profile Hidden Markov Models confirm to be one of the best ways to carry out this kind of work which also does not require a huge amount of computational power or time.

Figure 5: Multiple sequences alignment between the consensus sequence of the model (generated with hmmemit) and the best aligned domain of the false negative and false positive proteins. Cysteines with high information content are highlighted in red, other important residues in yellow.

5 References

1. Ranasinghe, S. & McManus, D. P. Structure and function of invertebrate Kunitz serine protease inhibitors. *Dev. Comp. Immunol.* **39**, 219–227 (2013).
2. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
3. Aprotin. *Drugbank* https://www.drugbank.ca/drugs/DB06692.
4. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10**, 402–415 (2009).
5. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
6. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
7. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
8. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268 (2004).
9. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
10. 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* **15**, 7 (2014).