Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Manuscript Category

A hidden Markov model for the functional annotation of kunitz-type domains

Matteo Bolner

University of Bologna

Abstract

Motivation: The in-silico discrimination of the presence of kunitz-type domains in a protein sequence is of critical importance for the research and development of pharmaceutical drugs; the aim of this project is the functional annotation of non-reviewed Uniprot sequences through a hidden Markov model built on the structural alignment of known kunitz-type domains.
Results: A hidden markov model able to discriminate between kunitz and non-kunitz domains was succesfully produced and optimized.
Supplementary information: Supplementary data are available at *https://github.com/matteobolner/laboratory-of-bioinformatics/tree/master/second_semester/project/* and in the attached file.

1 Introduction

The kunitz-type protein family [1] consists of small peptide sequences, usually between 50 and 60 residues long, which can exist as either standalone proteins or as domains of bigger proteins. They have a molecular weight of around 6 kDa, with an alpha+beta fold organization consisting of a two-stranded antiparallel β-sheet followed by an alpha helix; six cysteins form three disulphide bonds which stabilize the globular structure. Kunitz-type domains are mostly involved in the inhibition of protease enzymes, such as trypsin and kallikrein. In vertebrates, Kunitz-type proteins play a major role in inflammatory processes, while in invertebrates they are involved in a range of diverse functional roles, such as providing protection to parasitic organisms from host digestive protease enzymes, functioning as anti-coagulant factors, defending the organism against pathogens or functioning as toxins.[2] They are extensively studied for their possible applications, for example in the development of drugs like Aprotinin, a bovine pancreactic trypsin inhibitor (BPTI) used to reduce post-surgery bleeding and fibrinolysis, or Ecallantide, used for the treatment of angioedema.
The approach described in this article was used to generate a model of the kunitz-type protein family, which can be used to calculate whether a protein belongs to it or not by analyzing its sequence with respect to the model.

2 Methods

2.1 Datasets:

2.1.1 Model data

The profile HMM was built from a multiple sequence alignment of kunitz-type proteins using HMMER 3.2.1[3]; in order to obtain this multiple sequence alignment, the following steps were performed:

1. RCSB PDB [4] was browsed using the following filters:
   - PFAM ID PF00014;
   - Wild type protein, including expression tags and any percent coverage of Uniprot sequence;
   - Chain length between 40 and 70 residues;
   - X-ray resolution between 0 and 3.5 Å

   From the 159 resulting structures (including different chains), 2KNT was chosen as a good representative for the family.
2. In order to cross-validate the results, 2KNT was browsed through the pairwise alignment tool on PDBeFold [5] against the whole PDB and with the highest precision; the 322 results were compared to the PDB search results, and the 157 chains in common were selected.
3. The sequences obtained from the previous step were clustered using blastclust from the blast 2.2.26 package [6] in order to choose one representative for each cluster, so as not to build a biased model with structures too similar to each other. The sequence similarity was set to be at least 90%, while the coverage length was set to at least 80%.
4. From each one of the 16 clusters obtained, the entries with the lowest resolution were selected to build the model.
5. The 16 entries were aligned with PDBeFold's webtool for multiple comparison and 3D alignment of protein structures.

2.1.2 Optimization datasets

In order for the model to work as intended and produce useful results, it is necessary to determine a specific e-value treshold that maximizes its capacity to tell apart positive (kunitz) from negative (non kunitz) input sequences. To do so, two optimization datasets were built, both consisting of Uniprot/Swissprot [7] sequences.

- Positive set : all the reviewed sequences containing the kunitz-type domain PFam ID (PF00014); 322 sequences were obtained. From this set the sequences used to build the model were removed, in order to avoid redundancy and therefore biases in the optimization. Only 8 of the 16 model sequences were actually present with 100% sequence identity in the positive set after a BLASTP search against the database created with makeblastdb from the 322 sequences.

- Negative set: all the reviewed sequences of length between 40 and 500 residues, not containing the kunitz-type domain PFam ID (PF00014); 442444 sequences were obtained. The sequence length was limited to 40-500 residues in order not to skew the testing of the model, since sequences shorter than 40 residues can not contain the whole kunitz domain and sequences longer than 500 residues can contain it several times, or have a higher probability of containing sequences very similar to it.

2.2 Building and testing the model:

The multiple sequence alignment obtained from the structural alignment of 16 sequences at the end of section 2.1.1 was used as input for the building of the profile HMM, by using the hmmbuild program from HMMER. The model was tested against the two optimization sets with the hmmsearch program from HMMER; since hmmsearch doesnt return as output the sequences with an e-value deemed too high to be significant, for the negative set it was necessary to add the option to consider e-values up to 10^6. The positive set returned the expected 314 results, with corresponding e-value. Interestingly enough, the negative search returnet three sequences with an e-value low enough to be considered positive sequences (see Discussion section for more in-depth considerations). From the output of the positive set search the ID and e-value of every entry were extracted and labeled with the number 1; each e-value was normalized using the Bonferroni correction. The same was done for the negative set; however, most sequences in the negative set were not reported in the output of HMMSearch even after adding the option to consider e-values up to 10^6. This was solved by adding them manually to the output file and assigning them an e-value of 1.0, which is still too high for a sequence to be significant. Every e-value was then normalized with the Bonferroni correction, and 0 was added as a label. The two lists of e-value,id and label were merged in order to test the model performance.

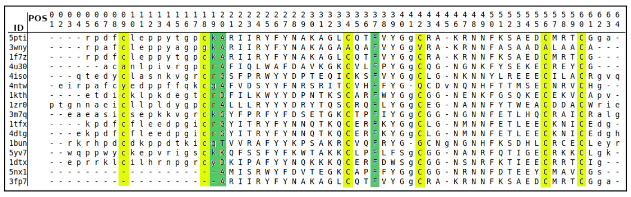
2.3 Evaluating the performance of the model:

In order to measure the performance of the model, a python script was devised(see supplementary materials); every unique e-value in the optimization set was used in an iteration of the script as the treshold with which to obtain the confusion matrix. From the confusion matrix the Matthews Correlation Coefficient (MCC) was calculated, along with the True Positive Ratio (TPR), False Positive Ratio (FPR) and Accuracy (ACC) (see supplementary materials). From the TPR and FPR, the Receiver Operating Characteristic (ROC) curve was plotted; the Area Under Curve (AUC) was also measured.

2.4 **Results and discussion**

2.4.1 Results

The model is consistent with the general description of a kunitz domain: as can be seen in the MSA (Figure 1), the six cysteins involved in the disulphide bonds are the most conserved residues in their respective positions; the same is for the active site residues, in position 19 and 20 in the model, which can be either an arginin(R) or a lysin(K) for position 19, and an alanin(A) or glycin(G) for position 17. Additionally, the phenilalanin residue in position 37 is highly conserved, along with other phenilalanines and tyrosines(Y) due to their role in the stabilization of the reactive site structure through internal hydrophobic interactions.



```
     POS 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6
ID       1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4
5pti     - - - - r p d f c l e p p y t g p c k A R I I R Y F Y N A K A G L C Q T F V Y G g C R A - K R N N F K S A E D C M R T C G g a -
3wny     - - - - r p a f c l e p p y a g p g k A R I I R Y F Y N A K A G A A Q A F V Y G g V R A - K R N N F A S A A D A L A A C A - - -
1f7z     - - - - r p d f c l e p p y t g p c k A R I I R Y F Y N A K A G L C Q T F V Y G g C R A - K R N N F K S A E D C M R T C G - - -
4u30     - - - - - - a c a n l p i v r g p c r A F I Q L W A F D A V K G K C V L F P Y G g C Q G - N G N K F Y S E K E C R E Y C G - - -
4iso     - - - q t e d y c l a s n k v g r c r G S F P R W Y Y D P T E Q I C K S F V Y G g C L G - N K N N Y L R E E E C I L A C R g v q
4ntw     - e i r p a f c y e d p p f f q k c g A F V D S Y Y F N R S R I T C V H F F Y G - Q C D v N Q N H F T T M S E C N R V C H g - -
1kth     - - - - e t d i c k l p k d e g t c r D F I L K W Y Y D P N T K S C A R F W Y G g C G G - N E N K F G S Q K E C E K V C A p v -
1zr0     p t g n n a e i c l l p l d y g p c r A L L L R Y Y Y D R Y T Q S C R Q F L Y G g C E G - N A N N F Y T W E A C D D A C W r i e
3m7q     - - e a e a s i c s e p k k v g r c k G Y F P R F Y F D S E T G K C T P F I Y G g C G G - N G N N F E T L H Q C R A I C R a l g
1tfx     - - - - k p d f c f l e e d p g i c r G Y I T R Y F Y N N Q T K Q C E R F K Y G g C L G - N M N N F E T L E E C K N I C E d g -
4dtg     - - - e k p d f c f l e e d p g i c r G Y I T R Y F Y N N Q T K Q C E R F K Y G g C L G - N M N N F E T L E E C K N I C E d g h
1bun     - - r k r h p d c d k p p d t k i c q T V V R A F Y Y K P S A K R C V Q F R Y G - G C N g N G N H F K S D H L C R C E C L e y r
5yv7     - - w q p p w y c k e p v r i g s c k K Q F S S F Y F K W T A K K C L P F L F S g C G G - N A N R F Q T I G E C R K K C L g k -
1dtx     - - e p r r k l c i l h r n p g r c y D K I P A F Y Y N Q K K K Q C E R F D W S g C G G - N S N R F K T I E E C R R T C I g - -
5nx1     - - - - - - - - - - - - - - - - - - - A M I S R W Y F D V T E G K C A P F F Y G g C G G - N R N N F D T E E Y C M A V C G s - -
3fp7     - - - - - - - - - - - - - - - - - - - A R I I R Y F Y N A K A G L C Q T F V Y G g C R A - K R N N F K S A E D C M R T C G g a -
```

Fig. 1. Multiple sequence alignment derived from structural alignment on PDBeFold, on the basis of which the HMM was built; in yellow are highlighted the conserved cysteins, while in green are highlighted the two conserved active site residues and the conserved phenylalanin.

Regarding the model performance (Table 1), the MCC closest to 1 corresponds to the optimal e-value treshold with which the model may be used in the testing of new sequences; from the data obtained in section 2.3 the highest MCC was 0.995, corresponding to an e-value of 6.37e-07.

Table 1. Summary of the model performance; T = True, P = Positive, F = False, N = Negative; ACC = Accuracy; TPR/FPR = True/False Positive Rate; MCC = Matthew's Correlation Coefficient

E-val treshold	TP	FP	FN	TN	ACC	TPR	FPR	MCC
1.083e-29	38	0	276	642	0.711	0.121	0.000	0.291
7.643e-26	167	1	147	641	0.845	0.532	0.002	0.654
6.369e-07	314	2	0	640	0.998	1.000	0.003	0.995

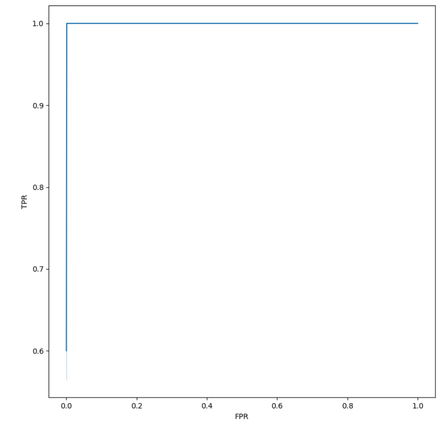
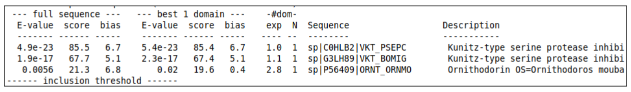


Fig. 2. ROC curve obtained from all the iterations of the model performance measurement; FPR indicates the False Positive Rate, while TPR indicates the True positive rate.

The ROC curve obtained (Figure 2) represents a near-ideal measure of separability: the TPR immediately climbs up to 1.0 while the FPR stays at around 0.0. The AUC was measured as 1.0, which implies that the model is very good at discriminating between positive and negative sequences.

2.4.2 Discussion

The three false positives obtained after testing the model with the optimal e-value treshold are described on Uniprot as Kunitz-Type serine protease inhibitors (C0HLB2, G3LH89) ,and BPTI/Kunitz inhibitor (P56409), but are not annotated with the PF00014 PFam ID (Figure 3). While the model considers them false positives because they are not labeled as positives, they appear to actually be Kunitz domain containing proteins. This is a further indication that the model is working as intended, and is able to discriminate positive from negative sequences; if nothing was known about these three sequences, they would be correctly classified.



```
 --- full sequence ---    --- best 1 domain ---    -#dom-
  E-value  score  bias    E-value  score  bias    exp  N  Sequence                 Description
  ------- ------ -----    ------- ------ -----    ---- --  --------                 -----------
  4.9e-23   85.5   6.7    5.4e-23   85.4   6.7     1.0  1  sp|C0HLB2|VKT_PSEPC       Kunitz-type serine protease inhibi
  1.9e-17   67.7   5.1    2.3e-17   67.4   5.1     1.1  1  sp|G3LH89|VKT_BOMIG       Kunitz-type serine protease inhibi
   0.0056   21.3   6.8       0.02   19.6   0.4     2.8  1  sp|P56409|ORNT_ORNMO      Ornithodorin OS=Ornithodoros mouba
------ inclusion threshold ------
```

Fig. 3. Output of HMMSearch for the three false positive sequences

In conclusion, we can say that our model is able to correctly classify a sequence as belonging or not to the Kunitz-Type protein family; it may therefore be used for functional annotation of unreviewed proteins.

References

[1] https://pfam.xfam.org/family/pf00014
[2] S.Ranasinghe, D.P.McManus; Structure and function of invertebrate Kunitz serine protease inhibitors, Developmental and Comparative Immunology 39 (2013) 219-227
[3] http://hmmer.org/
[4] https://www.rcsb.org/
[5] Protein structure comparison service PDBeFold at European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/ssm), authored by E. Krissinel and K. Henrick
[6] ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.26/
[7] https://www.uniprot.org/ -The Pfam protein families database in 2019: S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn Nucleic Acids Research (2019) doi: 10.1093/nar/gky995

-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.

-E. Krissinel and K. Henrick (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. D60, 2256—2268

-E. Krissinel and K. Henrick (2005). Multiple Alignment of Protein Structures in Three Dimensions. In: M.R. Berthold et.al. (Eds.): CompLife 2005, LNBI 3695, pp. 67–78. Springer-Verlag Berlin Heidelberg. -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.

-The UniProt Consortium UniProt: a worldwide hub of protein knowledge Nucleic Acids Res. 47: D506-515 (2019)