Sequence analysis May 31, 2023

Development of a Profile Hidden Markov Model for Accurate Identification of Kunitz-type Domains using Structural Alignment

Leonardo Politi[1]

[1]Master's Degree in Bioinformatics, University of Bologna, Italy

Abstract

Motivation: Protease inhibitors play major roles in infection, inflammation disorders and many other diseases. The BTPI-Kunitz domain, belonging to the trypsin inhibitors family, has garnered significant attention and has been extensively studied showing potential applications in medicine, agriculture, and biotechnology. Accurate classification and annotation of these domains is essential for understanding their function and designing targeted therapies. In this study two different Hidden Markov Models were built, based on a multiple structural alignment, to accurately identify the Kunitz-type domain in novel sequences. The performances of the two models were evaluated and compared over the entire UniProtKB/SwissProt database.

Results: Both models achieved very high performance scores, and they seems to be reliable and effective classifiers for the identification of the Kunitz domain. For the first model (mcc = 0.996, f1= 0.996, auc = 0,997) 1 false positive and 2 false negatives were obtained, while for the second one (mcc = 0.997, f1 = 0.997 , auc = 0.997) only 1 false negative was obtained.

Availability and implementation: All the datasets used for training and testing, the final hmm model and the python code are freely available at https://github.com/LeonardoPoliti/LB1-project-Kunitz-HMM

Contact: leonardo.politi2@studio.unibo.it

1 Introduction

Protease inhibitors are a group of proteins, which inhibit the function of proteases in the biological systems playing fundamental roles in cell biology. These proteins are involved in many processes, like cell proliferation and cell homeostasis[1]. Protease inhibitors are also involved in the processes of coagulation, fibrinolysis, and inflammatory process[3,13], including modulation of cytokine expression, signal transduction and tissue remodelling[4].

Among these proteases the most exhaustively studied mechanism of protein inhibitors is the standard mechanism of the serine proteases. They can be grouped in several families based on their sequence homology, position of the reactive site,

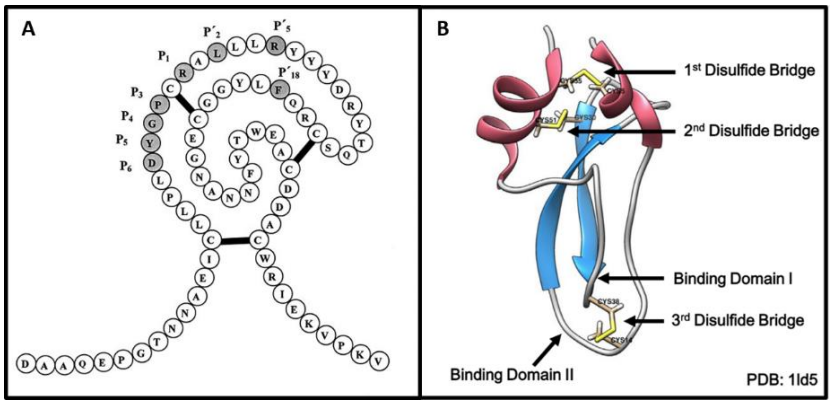


Figure 1: (A) Model structure of the first Kunitz-type domain (KD1) of human TFPI-2. (B) Classical Kunitz Inhibitor BPTI structure (pdb:1ld5). Conserved cysteine residues and the three disulphide bonds are highlighted in yellow, while the two antiparallel β-sheets are in blue and the α-helix in red.

structural characteristics and mechanism of action[8].

These inhibitors include the Kunitz type family, which contain a structurally conserved domain, typically composed of approximately 50-70 amino acids (6 kDa), stabilized by generally three disulfide bridges (*Fig. 1A*) [11]
They are particularly prevalent among metazoans, but they have been also identified in plants, microbes, and other organisms.
Kunitz domains are structurally characterized by two anti-parallel β-sheets, one or two helical regions and a solvent exposed loop (*Fig. 1B*), that determines the binding specificity[2].

Kunitz domains have been found to have additional functions, such as modulating ion channels or serving as toxins in venomous animals[10]. This further highlights their wide distribution and functional diversity across different organisms, and its evolutionary importance[9]. By studying the structure, function, and biological activities of the Kunitz domain, we can gain a deeper understanding of its mechanisms, interactions, and potential applications. This knowledge paves the way for the development of innovative therapies, diagnostic tools, and biotechnological solutions.

Given its importance, it's crucial to have effective tools contributing to the domain characterization. In this study, two hidden Markov models (HMMs) were constructed by aligning two different sets of proteins that contain the Kunitz domain. These alignments were generated using distinct approaches.
HMMs play a crucial role in capturing the probabilistic relationships between different positions in a sequence alignment. By considering both the observed residues and the hidden states representing various structural or functional motifs, HMMs can effectively model the sequence-structure relationship of a protein family. This provide a robust and probabilistic framework for accurately annotating new sequences, searching for additional homologs, and generating high-quality multiple sequence alignments[5,15].
The goal of this study was to generate a reliable classifier for the presence of the Kunitz domain in a sequence, and classify the entire SwissProt dataset. The performance of the two models were compared and the false positives and false negatives obtained at different E-value thresholds were manually evaluated.

2 Methods

2.1 Datasets preparation

To obtain the first training set, the RCSB PDB database was queried to fetch a collection of structures with annotated Kunitz domain by Pfam (Pfam AC: PF00014), resolution equal or below 3 Å and sequence length between 45 to 80 amino residues.

For the second training set, two pdb entry were selected as templates (pdb: 1bpi, 3tgi – uniprot: P00974) since their structures have been vastly characterized by numerous publication and they have very high resolution (1.09 Å, 1.80 Å) . To obtain the dataset the two structures were independently aligned against all the pdb dataset, using PDBeFold[7]. The search was done with 'high' precision and with a lowest acceptable match of 95%. Finally all the proteins obtained were merged in a single dataset and they were filtered with the same criterion of the first dataset: resolution equal or below 3 Å and sequence length between 45 to 80 amino residues.

To deal with redundancy, the two training sets obtained were clustered by 95% sequence identity (using the pdb advanced search) and only the representative protein (based on rmsd) was taken for each cluster. The final size of the training sets were: 27 sequences for the 'pfam' set and 34 sequences for the '1bpi+3tgi' set.

The test set used for both models consisted of two subsets. The first subset comprised reviewed proteins annotated by Pfam as containing the Kunitz domain, which served as the positive set. The second subset consisted of all other proteins that had not been annotated to contain the Kunitz domain, forming the negative set. Both models were evaluated on this common test set.

The positive test set was retrieved by querying the UniProtKB database for reviewed proteins with BPTI/Kunitz domain annotated by Pfam (pfam_id = PF00014). In this way 390 total sequences were found. From these, the proteins also selected for the training sets were removed resulting in 374 sequences for Model1 (pfam) and 379 for Model2 (1bpi+3tgi). Note that out of the first train set, only 16 samples were successfully mapped in SwissProt using their corresponding PDB identifiers. Similarly, in the second train set, only 11 samples were successfully mapped.

All the other SwissProt proteins that were not used for the positive and train sets formed the negative test set (569126 sequences).

The database searches described above were performed on May 24, 2023. Both training and test sets can be found in the supplemental material on GitHub.

2.2 Multiple structural alignments

PDBeFold was used to perform the multiple structural alignment for both training sets.
Some proteins have resulted to have multiple chains containing the Kunitz domain, in these cases a manual review of the chains was needed since selecting the first chain of each sequence resulted in a poor multiple alignment.

Train-set	RMSD	Q-score
pfam	0.7406	0.4376
1bpi+3tgi	0.8081	0.5985

Tab 1: scores of the multiple structural alignments for the two training sets used.

Due to the short length of the alignments, a manual trimming process was carried out, retaining only 59 positions. The remaining portion of the alignment mainly consisted of gaps, which could potentially introduce noise and adversely impact the accuracy of subsequent analyses. Notably, the deleted regions did not contain any cysteines, which are known to play a crucial role in maintaining the stability of the Kunitz domain[11].

2.3 Generation of the Profile HMMs

The Profile Hidden Markov Models were generated by running the 'hmmbuild' program of HMMER (v.3.3.2) with default options.

HMMER is a powerful software package specifically developed for constructing probabilistic models of protein and DNA sequence domain families[6]. 'Hmmbuild' reads the multiple structural alignment returned by PDBeFold in FASTA format and builds a new model.

The Skylign web tool allows the HMM's visualization via generation of a sequence logo[14]. (*Fig. 2*) In both generated models, it shows the presence of 6 highly conserved cysteine residues. This is consistent with crystallographic studies which generally show the presence of three stabilizing disulfide bridges[12]. It is possible to observe that the residues involved in the first and third disulfide bridge have significantly higher information content than the pair of the second disulphide bridge.

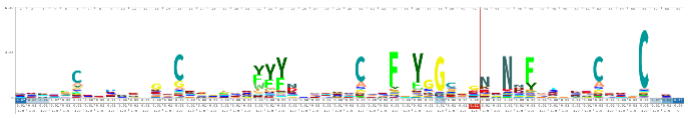


Figure 2: Sequence logo of a kunitz-hmm model obtained from Skyalign. **N.B:** This image is a place holder since the Syalign website stopped working for the last week and this is the only image I've done before. (it is not any of the models described in the paper!)

2.4 Model optimization and Performance evaluation

To conduct "threshold tuning" and determine the optimal E-value for classifying the test set, the test set of each model was shuffled and divided into two subsets (cv1 and cv2). Each subset was carefully constructed to contain an equal number of positive and negative examples, to ensure a balanced representation of the target class within each subset.

The models were matched with these subsets using the 'hmmsearch' program of HMMER

(v.3.3.2) with parameters: –Z 1 –max –noali. Since the main objective was to determine whether the sequences contained the Kunitz domain or not, only the E-value associated with the best matching domain was considered for further analysis and the E-value of proteins for which the domain was not identified, was manually set to 100.

Finally, the performance of the two models were evaluated over a range of e-values between 1 and $1e^{-20}$ for both subsets.
It is important to notice that the total test set comprised around 569,500 protein sequences, but only a small subset, approximately 380 sequences, constituted the positive set. This evident imbalance between the positive and negative classes highlights the limitation of using accuracy (AC) (1) as an evaluation metric. Instead, alternative approaches that are more resilient to skewed class distributions were employed to assess the model's performance.

$$(1) \qquad AC = \frac{TP + TN}{TP + FN + TN + FP}$$

A confusion matrix was constructed to compare the predicted labels against the actual class labels. In order to consider both precision and recall in the threshold selection process, multiple evaluation metrics were computed for each model. These metrics included the Matthews Correlation Coefficient (MCC) (2), F1 score (3), and Area Under the Curve (AUC).

$$(2) \qquad MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

$$(3) \qquad F1 = 2\frac{\left(\frac{TP}{TP+FP}\right) \cdot \left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}\right) + \left(\frac{TP}{TP+FP}\right)}$$

3 Results

To optimize the models and achieve an accurate identification of the presence of the Kunitz domain, the previously described performance analysis were performed for various E-value thresholds. Two test-subsets (cv1 and cv2) were initially used for the 'threshold-tuning', and both models obtain the highest score for the E-value $1e^{-08}$. This was then used for the final evaluation of the entire test set. (*Sup_Fig. 3-6*)

Model1 (mcc = 0.996, f1= 0.996, auc = 0,997) obtained 1 false positive and 2 false negatives (*Sup_Fig. 1*). Similarly, Model2 (mcc = 0.997, f1 = 0.997 ,auc = 0.997) obtained only 1 false negative (*Sup_Fig. 2*).

The two models showed highly comparable performances (*Fig. 3*). This suggest that both models are reliable and effective tools for accurately identify the presence of the Kunitz domain in diverse and novel protein sequences.

4 Discussion

Observing the classification results at different E-value thresholds *(Sup_Fig. 1,2)* we can notice that some of the negative examples have been assigned with a small E-value, resulting in relatively large number of false positives for thresholds lower than $1e^{-08}$. This follow the expectation, since the negative set was much bigger than the positive one.

To further investigated the low reliability of these classifications, the sequences misclassified in the threshold range $1e^{-05}$ to $1e^{-10}$ have been manually inspected.

In this range of E-values only four false negatives where identified for both models. D3GGZ8 and O62247 had shown serine protease activity in vitro but they lacks some catalytic features of serine proteases. The other two (P86963, Q11101) clearly contain the Kunitz domains but they have not been deeply studied, and they just have a slightly more ambiguous sequence.

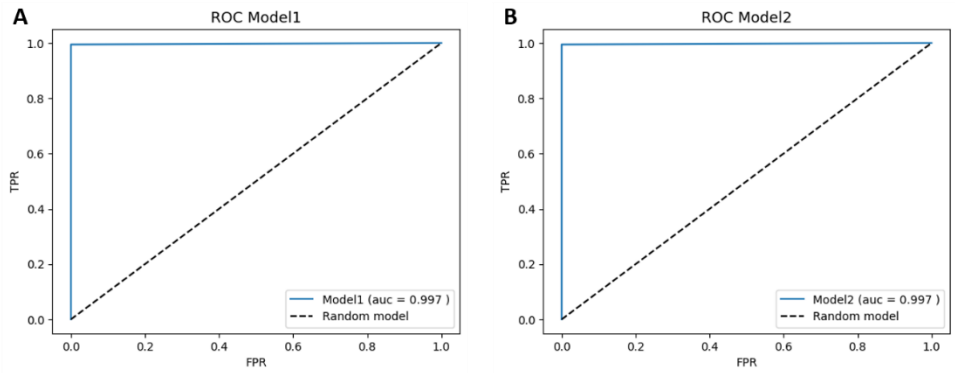


Figure 3: ROC curve plots of the final test sets for Model1 (**A**) and Model2 (**B**). The x-axis represents the False Positive Rate (FPR) and the y-axis the True Positive Rate (TPR)

The false positives sequences obtained are generally poorly annotated in Swissprot and don't have many related publications. Many of them, e.g, Q8WPG5 and Q09JW4, present the three typical disulfide bonds. Many others, like P84555 and P85039, are short fragments but they show some Kunitz key structural features (e.g., the conserved cysteines) and are likely pieces of a larger Kunitz domain.

Is important to notice that the majority of the false positives, have been in some ways annotated as Kunitz domains by other datasets, like InterPro and SUPFAM. Remember that the negative test set was composed by the proteins that have not been annotated with the kunitz-pfam identifier (PF00014), so these false positives are likely to had been misclassified or are not present in the pfam database.

To conclude, in this study the E-value threshold $1e^{-08}$ has shown the best performances for both models, but in reality it does not guarantee that it's the optimal choice in all scenarios, since the evaluation of performance can be influenced by the methods employed in selecting the testing set, which can introduce biases. To further enhance the annotation accuracy of the Pfam database, it is worthwhile to explore sequences that were misclassified by utilizing slightly different thresholds.

References

[1] Ascenzi, P., Bocedi, A., Bolognesi, M., Spallarossa, A., Coletta, M., De Cristofaro, R., & Menegatti, E. (2003). The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. *Current protein & peptide science*, *4*(3), 231–251.

[2] Chand, H. S., Schmidt, A. E., Bajaj, S. P., & Kisiel, W. (2004). Structure-function analysis of the reactive site in the first Kunitz-type domain of human tissue factor pathway inhibitor-2. *The Journal of biological chemistry*, *279*(17), 17500–17507.

[3] Choo, Y. M., Lee, K. S., Yoon, H. J., Qiu, Y., Wan, H., Sohn, M. R., Sohn, H. D., & Jin, B. R. (2012). Antifibrinolytic role of a bee venom serine protease inhibitor that acts as a plasmin inhibitor. *PloS one*, *7*(2), e32269.

[4] De Magalhães, M. T. Q., Mambelli, F. S., Santos, B. P. O., Morais, S. B., & Oliveira, S. C. (2018). Serine protease inhibitors containing a Kunitz domain: their role in modulation of host inflammatory responses and parasite survival. *Microbes and infection*, *20*(9-10), 606–609.

[5] Eddy S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755–763.

[6] Eddy S. R. (2011). Accelerated Profile HMM Searches. PLoS computational biology, 7(10), e1002195.

[7] E. Krissinel and K. Henrick (2003). Protein structure comparison in 3D based on secondary structure matching (PDBeFold) followed by Ca alignment, scored by a new structural similarity function. In: Andreas J. Kungl & Penelope J. Kungl (Eds.), Proceedings of the 5th International Conference on Molecular Structural Biology, Vienna, September 3-7, 2003, p.88.

[8] Laskowski, M., Jr, & Kato, I. (1980). Protein inhibitors of proteinases. *Annual review of biochemistry*, *49*, 593–626.

[9] Mishra M. (2020). Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. *Journal of molecular evolution*, *88*(7), 537–548.

[10] Mourão, C. B., & Schwartz, E. F. (2013). Protease inhibitors from marine venomous animals and their counterparts in terrestrial venomous animals. *Marine drugs*, *11*(6), 2069–2112.

[11] Ranasinghe, S., & McManus, D. P. (2013). Structure and function of invertebrate Kunitz serine protease inhibitors. Developmental and comparative immunology, 39(3), 219–227.

[12] Schwarz,H. et al. (1987) Stability studies on derivatives of the bovine pancreatic trypsin inhibitor. Biochemistry, 26, 3544–3551.

[13] Wan, H., Lee, K. S., Kim, B. Y., Zou, F. M., Yoon, H. J., Je, Y. H., Li, J., & Jin, B. R. (2013). A spider-derived Kunitz-type serine protease inhibitor that acts as a plasmin inhibitor and an elastase inhibitor. PloS one, 8(1), e53343.

[14] Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC bioinformatics, 15, 7.

[15] Yoon B. J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. Current genomics, 10(6), 402–415.

Supplementary Material:

threshold	mcc	f1	auc	accuracy	precision	recall	fp_ids	fn_ids
1e-05	0.981781	0.981627	0.999988	0.999975	0.963918	1	'Q8WPG5', 'P84555', 'P56409', 'Q09JW3', 'C5H8E7', 'Q09JW4', 'P0DJ63', 'P83605', 'P85039', 'P84556', 'P85040', 'P83604', 'P0DM47', 'P0DV02'	[]
1e-06	0.986819	0.986772	0.998655	0.999982	0.97644	0.997326	'Q8WPG5', 'P84555', 'P56409', 'Q09JW3', 'C5H8E7', 'Q09JW4', 'P0DJ63', 'P83605', 'P85039'	'D3GGZ8'
1e-07	0.989369	0.989362	0.997321	0.999986	0.984127	0.994652	'Q8WPG5', 'P84555', 'P56409', 'Q09JW3', 'C5H8E7', 'Q09JW4'	'O62247', 'D3GGZ8'
1e-08	0.995982	0.995984	0.997325	0.999995	0.997319	0.994652	'P84555'	'O62247', 'D3GGZ8'
1e-09	0.995979	0.995973	0.995989	0.999995	1	0.991979	[]	'P86963', 'O62247', 'D3GGZ8'
1e-10	0.994635	0.994624	0.994652	0.999993	1	0.989305	[]	'Q11101', 'P86963', 'O62247', 'D3GGZ8'

Sup_Figure 1: Final scores of Model1. The E-value threshold 1e-08 is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp_ids = uniprot ids of the false positives, fn_ids = uniprot ids of the false negatives.

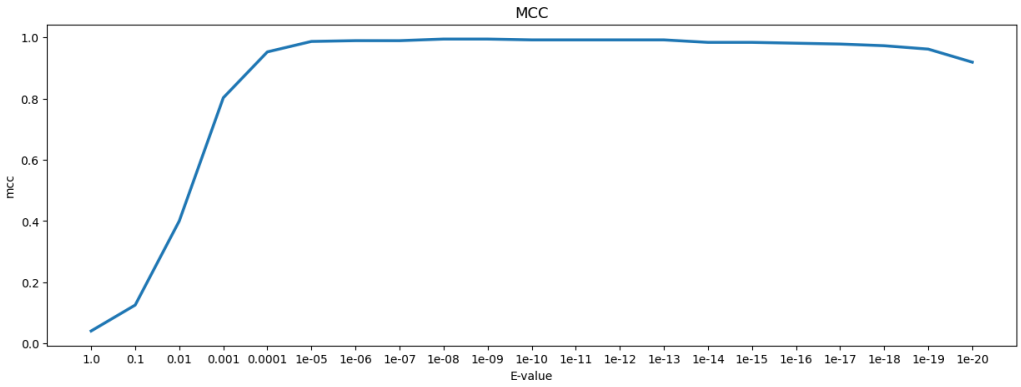
threshold	mcc	f1	auc	accuracy	precision	recall	fp_ids	fn_ids
1e-05	0.978177	0.978008	0.998667	0.99997	0.959391	0.997361	P84555', 'P85039', 'P0DJ63', 'P83605', 'Q09JW3', 'C5H8E7', 'P85040', 'Q09JW4', 'P84556', 'Q8WPG5', 'P56409', 'P71089', 'P0DM47', 'P36235', 'P0DV02', 'P83604'	D3GGZ8'
1e-06	0.983107	0.983051	0.997352	0.999977	0.971649	0.994723	P84555', 'P85039', 'P0DJ63', 'P83605', 'Q09JW3', 'C5H8E7', 'P85040', 'Q09JW4', 'P84556', 'Q8WPG5', 'P56409'	O62247', 'D3GGZ8'
1e-07	0.992103	0.992105	0.997358	0.999989	0.989501	0.994723	P84555', 'P85039', 'P0DJ63', 'P83605'	O62247', 'D3GGZ8'
1e-08	0.997356	0.997354	0.997361	0.999996	1	0.994723	[]	O62247', 'D3GGZ8'
1e-09	0.994705	0.994695	0.994723	0.999993	1	0.989446	[]	Q11101', 'P86963', 'O62247', 'D3GGZ8'
1e-10	0.994705	0.994695	0.994723	0.999993	1	0.989446	[]	Q11101', 'P86963', 'O62247', 'D3GGZ8'

Sup_Figure 2: Final scores of Model2. The E-value threshold 1e-08 is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp_ids = uniprot ids of the false positives, fn_ids = uniprot ids of the false negatives.

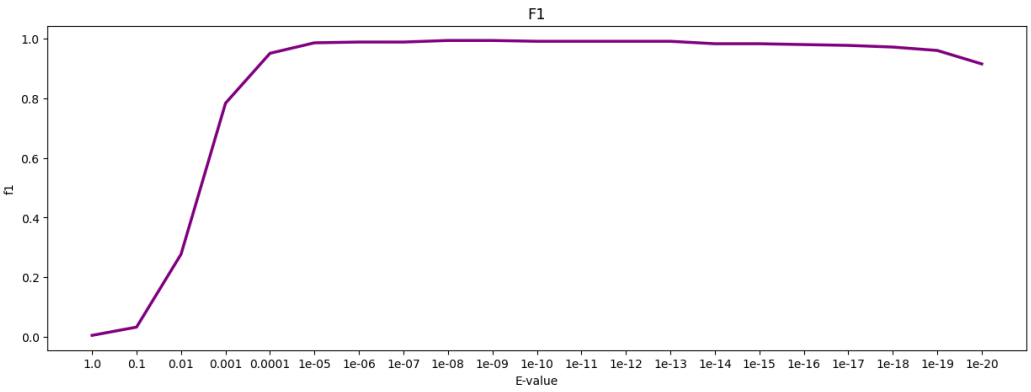
A

threshold	mcc	f1	auc	accuracy	fp	fn
1	0.041113	0.004681	0.860265	0.720713	79527	0
0.1	0.125752	0.032387	0.980366	0.960759	11174	0
0.01	0.400299	0.277037	0.998285	0.996572	976	0
0.001	0.802866	0.784067	0.999819	0.999638	103	0
0.0001	0.952736	0.951654	0.999967	0.999933	19	0
1e-05	0.986885	0.986807	0.999991	0.999982	5	0
1e-06	0.989466	0.989418	0.999993	0.999986	4	0
1e-07	0.989369	0.989362	0.997321	0.999986	3	1
1e-08	0.994649	0.994652	0.997324	0.999993	1	1
1e-09	0.994635	0.994624	0.994652	0.999993	0	2
1e-10	0.991941	0.991914	0.991979	0.999989	0	3
1e-11	0.991941	0.991914	0.991979	0.999989	0	3
1e-12	0.991941	0.991914	0.991979	0.999989	0	3
1e-13	0.991941	0.991914	0.991979	0.999989	0	3
1e-14	0.983816	0.983696	0.983957	0.999979	0	6
1e-15	0.983816	0.983696	0.983957	0.999979	0	6
1e-16	0.981093	0.980926	0.981283	0.999975	0	7
1e-17	0.978362	0.978142	0.97861	0.999972	0	8
1e-18	0.972878	0.972527	0.973262	0.999965	0	10
1e-19	0.961815	0.961111	0.962567	0.999951	0	14
1e-20	0.919148	0.915942	0.92246	0.999898	0	29

B



C

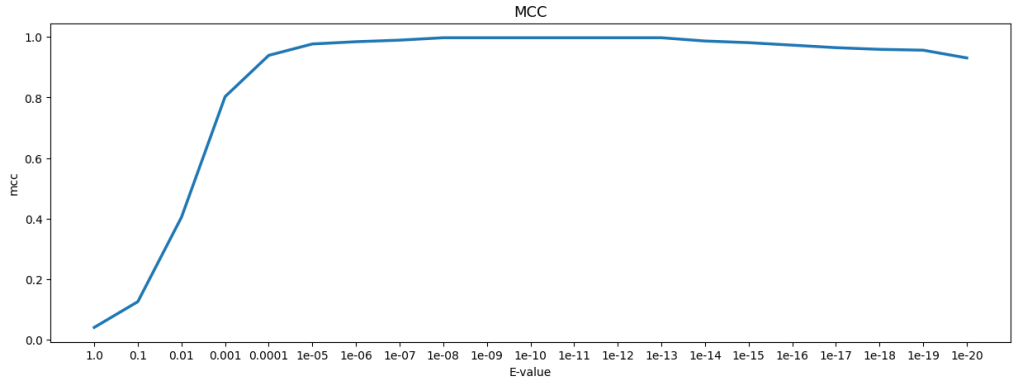


Sup_Figure 3: (**A**) Scores of Model1 on cv1 test set. The E-value threshold 1e-08 is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp = number false positives, fn = number of false negatives. (**B**,**C**) mcc and f1 plotted for all the tested E-value thresholds.

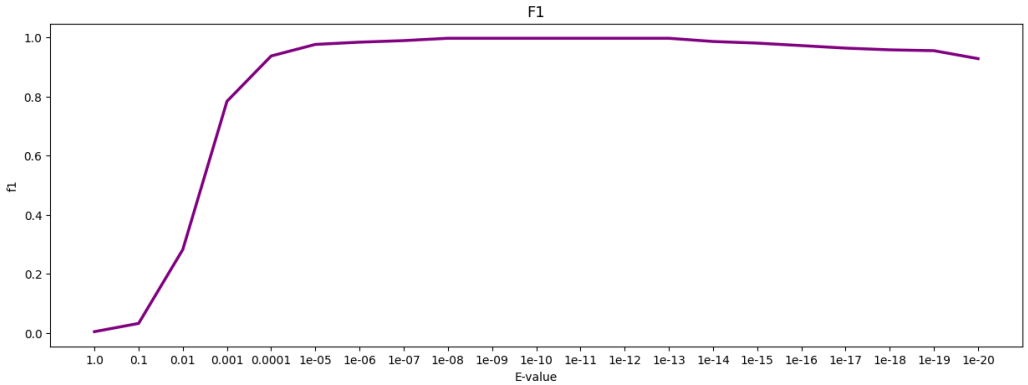
A

threshold	mcc	f1	auc	accuracy	fp	fn
1	0.041166	0.004689	0.860523	0.721229	79380	0
0.1	0.12618	0.032595	0.980496	0.961018	11100	0
0.01	0.404869	0.282477	0.998331	0.996664	950	0
0.001	0.802866	0.784067	0.999819	0.999638	103	0
0.0001	0.939147	0.937343	0.999956	0.999912	25	0
1e-05	0.976756	0.976501	0.999984	0.999968	9	0
1e-06	0.984172	0.984127	0.997317	0.999979	5	1
1e-07	0.989369	0.989362	0.997321	0.999986	3	1
1e-08	0.997321	0.997319	0.997326	0.999996	0	1
1e-09	0.997321	0.997319	0.997326	0.999996	0	1
1e-10	0.997321	0.997319	0.997326	0.999996	0	1
1e-11	0.997321	0.997319	0.997326	0.999996	0	1
1e-12	0.997321	0.997319	0.997326	0.999996	0	1
1e-13	0.997321	0.997319	0.997326	0.999996	0	1
1e-14	0.986532	0.98645	0.986631	0.999982	0	5
1e-15	0.981093	0.980926	0.981283	0.999975	0	7
1e-16	0.972878	0.972527	0.973262	0.999965	0	10
1e-17	0.964593	0.963989	0.965241	0.999954	0	13
1e-18	0.95903	0.958217	0.959893	0.999947	0	15
1e-19	0.956236	0.955307	0.957219	0.999944	0	16
1e-20	0.930717	0.928367	0.933155	0.999912	0	25

B



C

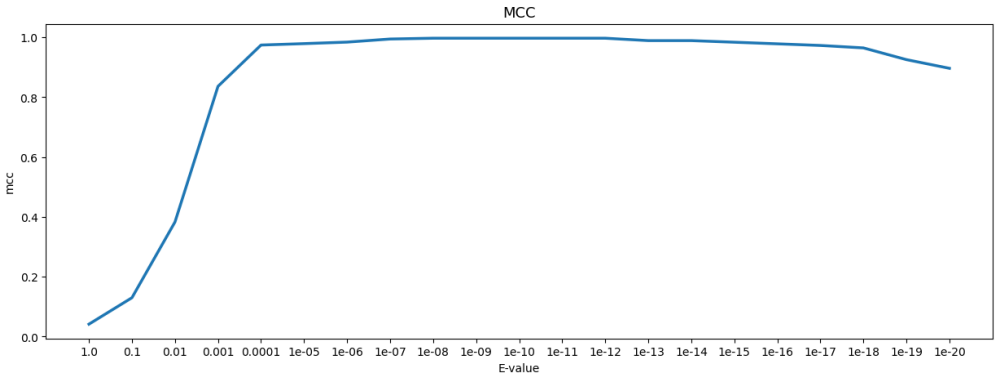


Sup_Figure 4: (**A**) Scores of Model1 on cv2 test set. The E-value thresholds range $1e^{-08}$- $1e^{-13}$ is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp = number false positives, fn = number of false negatives. (**B**,**C**) mcc and f1 plotted for all the tested E-value thresholds.

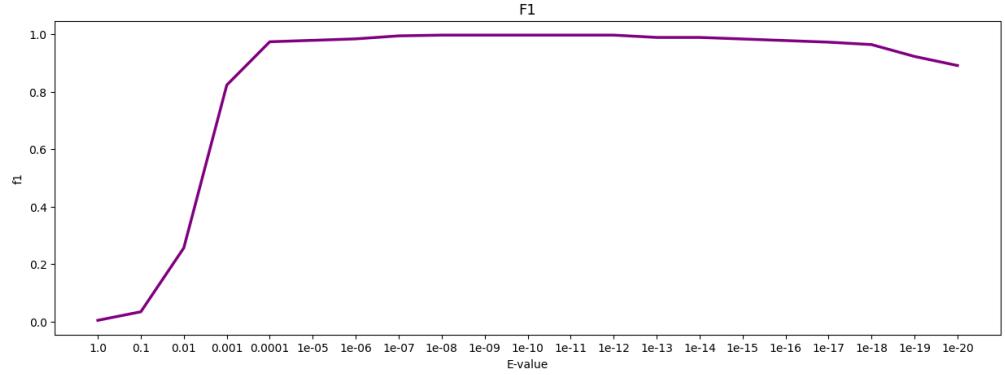
A

threshold	mcc	f1	auc	accuracy	fp	fn
1	0.041536	0.004764	0.861257	0.722699	78962	0
0.1	0.129948	0.034473	0.981398	0.96282	10587	0
0.01	0.383073	0.256793	0.998078	0.996158	1094	0
0.001	0.836541	0.823529	0.999858	0.999716	81	0
0.0001	0.974533	0.974227	0.999982	0.999965	10	0
1e-05	0.979272	0.979167	0.997342	0.999972	7	1
1e-06	0.984337	0.984293	0.997346	0.999979	5	1
1e-07	0.994705	0.994709	0.997353	0.999993	1	1
1e-08	0.997349	0.997347	0.997354	0.999996	0	1
1e-09	0.997349	0.997347	0.997354	0.999996	0	1
1e-10	0.997349	0.997347	0.997354	0.999996	0	1
1e-11	0.997349	0.997347	0.997354	0.999996	0	1
1e-12	0.997349	0.997347	0.997354	0.999996	0	1
1e-13	0.989354	0.989305	0.989418	0.999986	0	4
1e-14	0.989354	0.989305	0.989418	0.999986	0	4
1e-15	0.983989	0.983871	0.984127	0.999979	0	6
1e-16	0.978593	0.978378	0.978836	0.999972	0	8
1e-17	0.973168	0.972826	0.973545	0.999965	0	10
1e-18	0.964974	0.964384	0.965608	0.999954	0	13
1e-19	0.925776	0.923077	0.928571	0.999905	0	27
1e-20	0.896732	0.891496	0.902116	0.99987	0	37

B



C

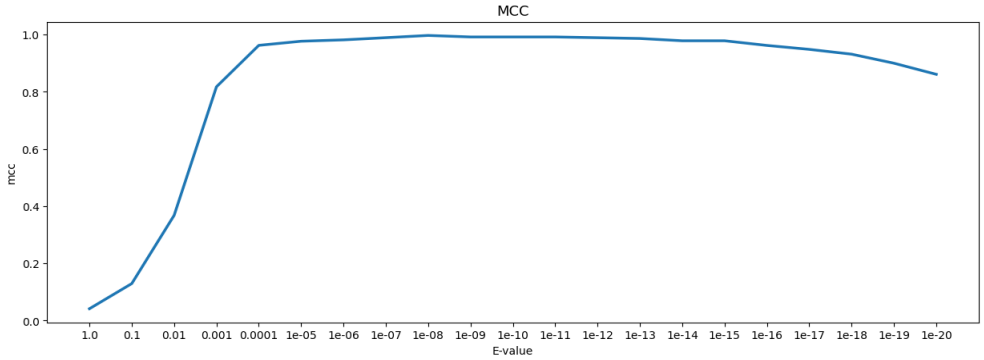


Sup_Figure 5: (**A**) Scores of Model2 on cv1 test set. The E-value threshold 1e-08 is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp = number false positives, fn = number of false negatives. (**B**,**C**) mcc and f1 plotted for all the tested E-value thresholds.

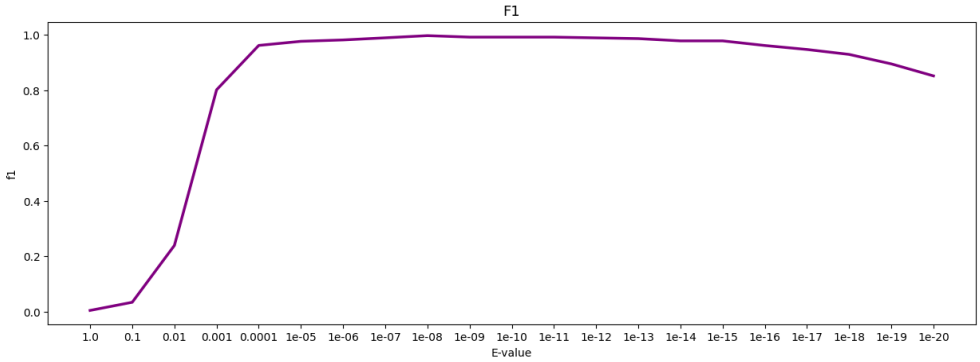
A

threshold	mcc	f1	auc	accuracy	fp	fn
1	0.041564	0.004776	0.860862	0.72191	79187	0
0.1	0.129568	0.03429	0.981196	0.962417	10702	0
0.01	0.368139	0.239596	0.997881	0.995765	1206	0
0.001	0.817798	0.801688	0.999835	0.99967	94	0
0.0001	0.962694	0.962025	0.999974	0.999947	15	0
1e-05	0.97711	0.976864	0.999984	0.999968	9	0
1e-06	0.981889	0.981818	0.997358	0.999975	6	1
1e-07	0.989535	0.989529	0.997363	0.999986	3	1
1e-08	0.997363	0.997361	0.997368	0.999996	0	1
1e-09	0.992069	0.992042	0.992105	0.999989	0	3
1e-10	0.992069	0.992042	0.992105	0.999989	0	3
1e-11	0.992069	0.992042	0.992105	0.999989	0	3
1e-12	0.989411	0.989362	0.989474	0.999986	0	4
1e-13	0.986746	0.986667	0.986842	0.999982	0	5
1e-14	0.978707	0.978495	0.978947	0.999972	0	8
1e-15	0.978707	0.978495	0.978947	0.999972	0	8
1e-16	0.962429	0.961749	0.963158	0.999951	0	14
1e-17	0.948652	0.947368	0.95	0.999933	0	19
1e-18	0.93185	0.929577	0.934211	0.999912	0	25
1e-19	0.900235	0.895349	0.905263	0.999874	0	36
1e-20	0.861381	0.851964	0.871053	0.999828	0	49

B



C



Sup_Figure 6: (**A**) Scores of Model2 on cv2 test set. The E-value threshold 1e-08 is highlighted since it maximizes the performance of the classifier. Mcc = Matthews Correlation Coefficient, auc = Area Under the Curve, fp = number false positives, fn = number of false negatives. (**B**,**C**) mcc and f1 plotted for all the tested E-value thresholds.