Project Report

Annotation of the Kunitz Domain using Hidden Markov Models

Roncelli Stefano

International Master in Bioinformatics, Univerisity of Bologna, Bologna

Latest revision Sunday 14th June, 2020

Abstract

Motivation: Hidden Markov Models are efficient tools for the analysis of domains. This work aims to briefly discuss their construction and application for one of the most studied protein domains: the Kunitz/BPTI domain.

Results: Once again, hidden Markov models proved to be very efficient in the classification of protein sequences. The model built was able to label correctly all the proteins with virtually no error. When dealing with protein sequences of unknown function, hidden Markov models can be a useful asset to improve their annotation.

Contact: stefano.roncelli@studio.unibo.it

Supplementary information: Available at https://github.com/saiden89/kunitz-project.

1 Introduction

The Kunitz/Bovine Pancreatic Trypsin Inhibitor (BPTI) domain is one of the most studied protein folds in structural biology1. Discovered nearly a century ago (Kraut *et al.*, 1930; Kunitz and Northrop, 1936) it was labeled as a selective inhibitor for serine protease activity. Through the years however many proteins with different function were shown to belong to the Kunitz/BPTI family, despite having no protease inhibition capabilities whatsoever. Indeed, many Kunitz-like snake toxins act as ion channel modulator, and some recent studies have suggested its involvement in the cuticle formation of some nematodes (Page *et al.*, 2006; Stepek *et al.*, 2010). It is a rather short (50-60 aa) globular domain, capable of exercising its function even as an independent unit. These features, along with its affinity for plasmin, made possible its application as an anti-hemorrhagic drug. Regardless of their function, all Kunitz-like proteins share the same core. The classical arrangement of the secondary elements consists of a 3_{10} helix near the N-terminus, two antiparallel β-strands and an α-helix beside the C-terminus. The loops connecting these elements usually show a certain degree of flexibility both in spatial position and length, especially in regard to Kunitz proteins that deviate from the classical inhibition function. Sequencewise, one of the most striking features is the presence of six highly conserved cysteines, which are involved in the stabilization of the tertiary structure through the formation of three disulphide bridges. Therefore, it should come to no surprise that they are the most conserved elements. This work aims to provide a basic workflow for the construction of a hidden Markov Model that is able to discriminate sequences harboring a Kunitz domain from those that do not. The process starts from the critical selection of the structures that will be used for the construction of the HMM, which are referred to as seeds. After this part, a benchmarking validation is used to provide an evaluation of the model performance. The results are then critically assessed in relation to the information currently available on the topic.

2 Methods

2.1 Seed curation and construction of the HMM

For the construction of the seed alignment, the UniprotKB (The UniProt Consortium, 2019) - SwissProt database (as of April_2020) was queried to retrieve all the sequences with an associated BPTI/Kunitz domain as in Pfam (El-Gebali *et al.*, 2019) (PF00014), limiting the search to the ones with a corresponding structure in PDB (Berman, 2000). To have the best possible quality for the seed alignment, the following criteria were applied in the subsequent filtering step:

- Actual presence of the Kunitz domain in the deposited structure
- Perfect coverage of the domain within the structure
- Highest possible resolution
- Absence of mutations
- X-ray diffraction structures only

Even though this manual approach should avoid redundancy altogether, clusterization of the seed sequences was performed with CD-HIT v4.8.1 (Li and Godzik, 2006) with default parameters. The structures that passed the selection were retrieved in order to perform multiple structural alignment via CATH-SUPERPOSE v0.16.5 with the intention to retrieve a high quality multiple sequence alignment. CATH-SUPERPOSE implements the CATH-SSAP (Taylor and Orengo, 1989) algorithm which performs the all-vs-all pairwise structural alignments. CATH-SSAP is particularly well suited algorithm to superpose structures that share a common motif or domain, as in our case. It is worth noting that this procedure emphasizes maximum overlap at the core level even at the expense of the overall RMSD, which may be higher than one obtained by more traditional structural aligners. Nevertheless, since we are only interested in the sequence alignment within the boundaries of the Kunitz domain, this caveat will not impact any result. The produced multiple sequence alignment was trimmed in order to include only the residues belonging to the domain and subsequently used as input for HMMER v3.3 (hmmer.org) `hmmbuild` for the retrieval of the HMM for the domain.

2.2 Model Validation

2.2.1 Construction of the datasets

The positive dataset for the model validation was constructed by retrieving from Swiss-Prot all the proteins ID with an annotated PFAM BPTI/Kunitz domain, without including the sequences used in the seed alignment. The negative dataset is made up of all the proteins without an annotated PFAM Kunitz domain. Both positives and negatives were merged in an unique dataset with 342 and 561894 entries respectively.

2.2.2 Using the profile against the full dataset

A well-performing HMM should be able to report Kunitz bearing sequences with low E-values and proteins devoid of such domain with higher E-values. This reasoning was applied with HMMER's `hmmsearch` tool used against the prepared dataset. All the heuristics filters were turned off (`--max` parameter); moreover, for normalization purposes the Initial search space (`Z`) was set to 1. Since the default inclusion E-value threshold of 10 prevents low scoring hits being reported, all the entries not included in the output were reintroduced with an arbitrary E-value equal to such threshold and labeled as negatives accordingly. For every hit both the global and the best domain E-values were kept for benchmarking. The retrieval and manipulation of the FASTA files and HMMER outputs were done using Biopython (Cock *et al.*, 2009) and then handled as Pandas (Hawkins, 2020) objects.

2.2.3 Validation of the profile

Validation of the profile was accomplished with a 5-fold stratified cross validation of the whole dataset, implemented via the `StratifiedKFold` function in SciKit Learn v0.23.1 (Pedregosa *et al.*, 2011) with `random_state` set to 42. Global and best-domain E-values constitute the feature on which the estimator will base the prediction; the true class of every entity represents the dependent variable. Two separate runs were performed for each feature. The entries for each of the five training sets were scanned by a customized Python pipeline to retrieve the E-value threshold that maximizes the Matthews Correlation Coefficient (MCC)(equation 2). The accuracy (ACC) was also computed as a performance metric, but never used for the training of the classifier (equaiton 1). The use of the MCC instead of ACC is motivated by the pronounced class imbalance of positives and negatives, with the negatives being several orders of magnitude more abundant than the positive. Indeed, the MCC is more resilient to class imbalance, and thefore chosen as the reference performance metrics. The search was conducted on E-values from 10^{-20} to 5 with a logarithmic step increase of 0.5. If multiple E-values provided the same MCC, then the average between those values was used as threshold. For each training test the newfound threshold was used as a categorical discriminant to predict the class of each entry in the test set as either Kunitz or non-Kunitz. Entries that fell below the threshold were labeled as predicted positive, the ones above were labeled as negative. All the mismatches for each of the test folds were manually analyzed to assess the reasons behind the misclassification. Finally, a comprehensive E-value was computed averaging the best thresholds found in the five training instances for both global and best domain datasets.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

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

where:

TP is the number of true positives

TN is the number of true negatives

FP is the number of false positives

FN is the number of false negatives

3 Results

3.1 Seed curation and construction of the HMM

Only thirty entries in UniProt with a Kunitz/BPTI domain have a corresponding entry in the PDB, making it possible to apply manual curation in the selection of the seeds. Just seventeen of these entries were retained; indeed, thirteen were discarded since they failed to meet the inclusion criteria. P12023, Q06481. Q7TQN3, albeit having a structure deposited in the PDB, were lacking the Kunitz domain. P08592 had only fragmented structures, which are not suitable for our purposes. P84875 had no wild type structures. Q5ZPJ7, C0HLB2, P68425, P0DJ46, P25660, P00979, P00981, Q96NZ8 had only NMR resolved structures. The retained atomic models were selected applying a best resolution preference rule. A summary of the seed alignment construction process is summarized in Table S1. The clusterization performed by CD-HIT grouped all the seed candidates in their own cluster, which means there was no redundancy in the seeds.

3.2 Superposition and multiple sequence alignment

The CATH-SSAP algorithm yielded a very good superposition at the level of the core (Figure 1), from which the sequence alignment was inferred (Figure 2). All the Kunitz domain of all the seed structures were correctly aligned, with no discernable difference in performance in respect to a stricly structural alignment method. The output was modified by a custom script retaining only the residues belonging to the domain, ensuring minimal noise for the generation of the HMM profile. From this very high-quality alignment, HMMER was able to reconstruct a profile nearly identical to the one retrievable from Pfam (Figure 3). Most importantly, all the conserved key positions, such as the disulfide-bonded cysteines were retained wherever possible.

3.3 Validation of the profile

The `hmmsearch` against the full dataset exhibits a marked class separation between positives and negatives (Figure 4). Consequently, the 5-fold cross validation procedure reported very few mismatches with no difference in classification among the E-values of the best domain and hit. The average thresholds for the datasets were $1.43 \cdot 10^{-9}$ for the global and $2.54 \cdot 10^{-9}$ for the best domain (Table 1). In both instances the false negatives are: Q11101, D3GGZ8, O62247 whereas the only false positive is G3LH89. A detalied view on the performances on each of the folds is reported in Table S2.

Table 1. Summary of the performance for both datasets.

Dataset	Avg. E-value	Avg MCC	Avg Accuracy
Global	$1.433\,87 \cdot 10^{-9}$	$0.994\,15$	$0.999\,99$
Best domain	$2.541\,07 \cdot 10^{-9}$	$0.994\,15$	$0.999\,99$

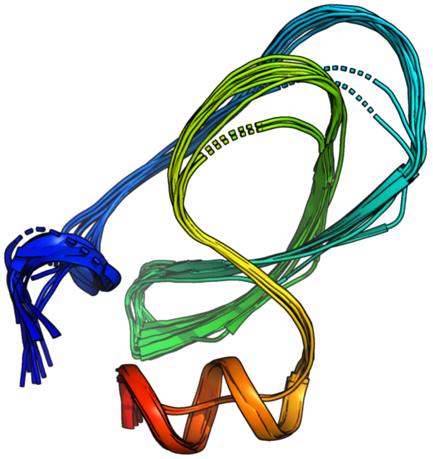
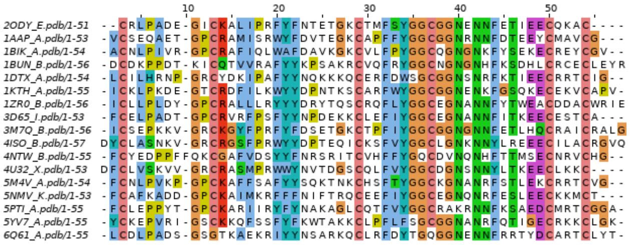


Fig. 1. Superposition of the seed structures at the core level. Image obtained with PyMol (Schrödinger, LLC, 2015).



```
                             10        20        30        40        50
2ODY_E.pdb/1-51     --CRLPADE-GICKALIPRFYFNTETGKCTMFSYGGCGGNENNFETIEECQKAC----
1AAP_A.pdb/1-53     -VCSEQAET-GPCRAMISRWYFDVTEGKCAPFFYGGCGGNRNNFDTEEYCMAVCG---
1BIK_A.pdb/1-54     -ACNLPIVR-GPCRAFIQLWAFDAVKGKCVLFPYGGCQGNGNKFYSEKECREYCGV--
1BUN_B.pdb/1-56     -DCDKPPDT-KICQTVVRAFYYKPSAKRCVQFRYGGCNGNGNHFKSDHLCRCECLEYR
1DTX_A.pdb/1-54     -LCILHRNP-GRCYDKIPAFYYNQKKKQCERFDWSGCGGNSNRFKTIEECRRTCIG--
1KTH_A.pdb/1-55     -ICKLPKDE-GTCRDFILKWYYDPNTKSCARFWYGGCGGNENKFGSQKECEKVCAPV-
1ZR0_B.pdb/1-56     -ICLLPLDY-GPCRALLLRYYYDRYTQSCRQFLYGGCEGNANNFYTWEACDDACWRIE
3D65_I.pdb/1-53     -FCELPADT-GPCRVRFPSFYYNPDEKKCLEFIYGGCEGNANNFITKEECESTCA---
3M7Q_B.pdb/1-56     -ICSEPKKV-GRCKGYFPRFYFDSETGKCTPFIYGGCGGNGNNFETLHQCRAICRALG
4ISO_B.pdb/1-57     DYCLASNKV-GRCRGSFPRWYYDPTEQICKSFVYGGCLGNKNNYLREEECILACRGVQ
4NTW_B.pdb/1-55     -FCYEDPPFFQKCGAFVDSYYFNRSRITCVHFFYGQCDVNQNHFTTMSECNRVCHG--
4U32_X.pdb/1-53     DFCLVSKVV-GRCRASMPRWWYNVTDGSCQLFVYGGCDGNSNNYLTKEECLKKC----
5M4V_A.pdb/1-54     -FCNLPVKP-GPCKAFFSAFYYSQKTNKCHSFTYGGCKGNANRFSTLEKCRRTCVG--
5NMV_K.pdb/1-53     -FCAFKADD-GPCKAIMKRFFFNIFTRQCEEFIYGGCEGNQNRFESLEECKKMCT---
5PTI_A.pdb/1-55     -FCLEPPYT-GPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA-
5YV7_A.pdb/1-55     -YCKEPVRI-GSCKKQFSSFYFKWTAKKCLPFLFSGCGGNANRFQTIGECRKKCLGK-
6Q61_A.pdb/1-55     -LCDLPADS-GSGTKAEKRIYYNSARKQCLRFDYTGQGGNENNFRRTYDCARTCLYT-
```

Fig. 2. Multiples sequence alignment of the seeds. Image obatined with Jalview 2 (Waterhouse et al., 2009)

4 Discussion

The manual curation of the seed alignment was made possible by the limited amount of proteins that satisfied the initial requirements (i.e. have any structure and a Kunitz domain). Such approach would not have been possible if there were a more substantial quantity of entries to begin with, at least with this very workflow. A possible solution to tackle a wider search space for the seeds may be an iterative approach, in which an initial, manually curated, seed is used to identify the most related sequences. Such procedure can be done iteratively until most of the search space has been scanned, to account for the natural diversity of the sequences that make up the domain. Nevertheless, the approach described here demonstrated how even a fistful of high-quality structures can be used in building a HMM that is effectively a representative of the domain. This was possible also because of the nature of the domain in study; indeed, the conservation of very important residues (i.e. the cysteines) necessary to achieve a proper fold plays favorably to our advantage. The outstanding capabilities of HMMs in domain pattern recognition are well reflected also in our study. The quality of the seeds made possible to build a model with virtually no noise, which in turn led to a strong class separation in the screening phase. With these premises, the subsequent cross-validation procedure so trivial that it would have been possible to perform it even by hand. The basic algorithm involved in the training successfully placed the E-value cutoff threshold in each of the five subsets. The misclassification of G3LH89 has not really to be taken into account, since it was wrongly labeled to begin with. The criteria to label as positive only the entries with an associated Kunitz Pfam domain backfired, since many other databases (e.g. Gene3D, SUPFAM, PROSITE) rightfully place it among the other Kunitz domain bearing proteins. Two false negatives (D3GGZ8 and O62247) are borderline cases. In fact, even though they are thought to belong to the Kunitz family, their serine-protease inhibition capabilities are disputed, as they are though to be involved in the cuticle formation pathway (Page *et al.*, 2006; Stepek *et al.*, 2010). Finally, Q11101 has no controversial aspects beside its poor annotation. It would be interesting to see in the future how the improvement annotations, alongside the increased availability of structures will reflect on the performance of this kind of classifiers. As for now, with the data at hand we can hardly question their efficiency.

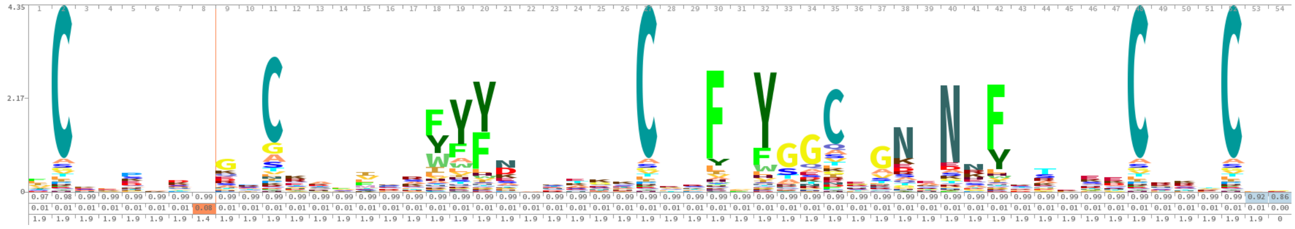


Fig. 3. HMM logo of the BPTI/Kunitz domain. Generated by Skilign (Wheeler et al., 2014).

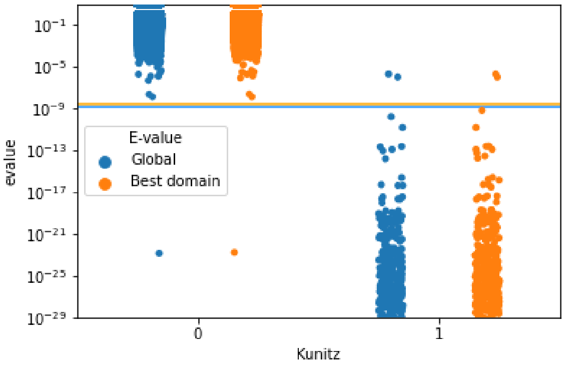


Fig. 4. Scatterplot of the entire dataset. E-values smaller that 10^{-20} are omitted for clarity purposes. The two horizontal lines are the average of the E-value thresholds for both datasets. Image obtained with Seaborn (Hunter, 2007).

References

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

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, **47**(D1), D427–D432.

Hawkins, S. (2020). Pandas-dev/pandas: Pandas 1.0.4. Zenodo.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, **9**(3), 90–95.

Kraut, H., Frey, E., and Werle, E. (1930). über die Inaktivierung des Kallikreins.(VI. Mitteilung Über dieses Kreislaufhormon.). *Hoppe-Seyler\' s Zeitschrift für physiologische Chemie*, **192**(1-3), 1–21.

Kunitz, M. and Northrop, J. H. (1936). ISOLATION FROM BEEF PANCREAS OF CRYSTALLINE TRYPSINOGEN, TRYPSIN, A TRYPSIN INHIBITOR, AND AN INHIBITOR-TRYPSIN COMPOUND. *The Journal of General Physiology*, **19**(6), 991–1007.

Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Page, A. P., McCormack, G., and Birnie, A. J. (2006). Biosynthesis and enzymology of the Caenorhabditis elegans cuticle: Identification and characterization of a novel serine protease inhibitor. *International Journal for Parasitology*, **36**(6), 681–689.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**(85), 2825–2830.

Schrödinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

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), 1–11.

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology*, **208**(1), 1–22.

The UniProt Consortium (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.

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), 7.

Supplementary material

Table S1. Summary of the seed candidates

Uniprot ID	Representative PDB ID	UniProt Name	Experimental Method	Exclusion reason
P05067	1AAP:A	A4_HUMAN	XRC	
P02760	1BIK:A	AMBP_HUMAN	XRC	
Q8WPI2	2ODY:E	BOOH2_RHIMP	XRC	
P00974	5PTI:A	BPT1_BOVIN	XRC	
P12111	1KTH:A	CO6A3_HUMAN	XRC	
A0A1Z0YU59	5M4V:A	MAMB1_DENAN	XRC	
O43278	4ISO:B	SPIT1_HUMAN	XRC	
O43291	4U32:X	SPIT2_HUMAN	XRC	
P10646	5NMV:K	TFPI1_HUMAN	XRC	
P48307	1ZR0:B	TFPI2_HUMAN	XRC	
Q90WA1	3D65:I	VKT1_PSETT	XRC	
P31713	3M7Q:B	VKT1_STIHL	XRC	
G9I929	4NTW:B	VKTA_MICTN	XRC	
P00989	1BUN:B	VKTH2_BUNMU	XRC	
P00980	1DTX:A	VKTHA_DENAN	XRC	
P81658	5YV7:A	VKTHC_DENAN	XRC	
P0C1X2	6Q61:A	VKTS1_CONST	XRC	
Q5ZPJ7		VKT_NAJAT		Only NMR structure available
C0HLB2		VKT_PSEPC		Only NMR structure available
P68425		VKT1A_CYRSC		Only NMR structure available
P0DJ46		VKT21_LYCMC		Only NMR structure available
P25660		VKT9_BUNFA		Only NMR structure available
P00979		VKTH1_DENPO		Only NMR structure available
P00981		VKTHK_DENPO		Only NMR structure available
Q96NZ8		WFKN1_HUMAN		Only NMR structure available
P12023		A4_MOUSE		No structure available
P08592		A4_RAT		Only fragmented structures
Q06481		APLP2_HUMAN		No structure available
P84875		PCPI_SABMA		Only mutant structures
Q7TQN3		WFKN2_MOUSE		No structure available

Table S2. Detailed performance of the classifier for global E-value and best domain E-value datasets.

Sample	True negatives	True positives	False positives	False negatives	Selected E-value threshold	Train Accuracy	Train MCC	Test Accuracy	Test MCC
Fold 1 - Global	112379	69	0	0	$1.778\,28 \cdot 10^{-9}$	0.999 99	0.994 49	1.000 00	1.000 00
Fold 2 - Global	112379	67	0	2	$5.623\,41 \cdot 10^{-11}$	1.000 00	0.996 33	0.999 98	0.985 39
Fold 3 - Global	112378	68	0	0	$1.778\,28 \cdot 10^{-9}$	0.999 99	0.994 51	1.000 00	1.000 00
Fold 4 - Global	112379	67	0	1	$1.778\,28 \cdot 10^{-9}$	1.000 00	0.996 35	0.999 99	0.992 62
Fold 5 - Global	112379	68	1	0	$1.778\,28 \cdot 10^{-9}$	1.000 00	0.996 34	0.999 99	0.992 72
Fold 1 - Best domain	112379	69	0	0	$1.778\,28 \cdot 10^{-9}$	0.999 99	0.994 49	1.000 00	1.000 00
Fold 2 - Best domain	112379	67	0	2	$5.623\,41 \cdot 10^{-11}$	1.000 00	0.996 33	0.999 98	0.985 39
Fold 3 - Best domain	112378	68	0	0	$1.778\,28 \cdot 10^{-9}$	0.999 99	0.994 51	1.000 00	1.000 00
Fold 4 - Best domain	112379	67	0	1	$1.778\,28 \cdot 10^{-9}$	1.000 00	0.996 35	0.999 99	0.992 62
Fold 5 - Best domain	112379	68	1	0	$1.778\,28 \cdot 10^{-9}$	1.000 00	0.996 34	0.999 99	0.992 72