Building a Profile Hidden Markov Model for Identifying the Bovine Pancreatic Trypsin Inhibitor Domain

Immanuela Antigone Engländer

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum Università di Bologna
Address Via San Donato 19/2, Bologna, Italy

*To whom correspondence should be addressed.

Received on 14.06.2020

Abstract

Motivation: The advent of both high-throughput sequencing techniques and high-throughput proteomics made it possible to generate astronomic amounts of data. One of today's challenges is to generate tools that enable fast yet reliable recognition of domains. Such tools must facilitate the classification of specific domains, helping us to endow the protein with structural and functional features and ultimately placing it into a given family. This investigation aimed at developing a profile Hidden Markov Model able to identify the bovine pancreatic trypsin inhibitor domain. It is a Kunitz type domain and a protease inhibitor. This globular protein has a length in the range of 50-70 amino acids and has found to be effective in the treatment of hyperfibrinolysis. This finding has lead to a spike in interest and extensive studies by pharmaceutical companies.

Results: The profile hidden Markov model achieved to classify the Kunitz-type proteins at a best threshold of 10^{-8} with an accuracy of 0.999 and a Matthews correlation coefficient of 0.994. Out of 178 positives two false negatives where classified at this threshold.

Contact: immanuela.englander@studio.unibo.it

Supplementary information: Supplementary data are available on *github*.

1 Introduction

The *bovine pancreatic trypsin inhibitor* (BPTI). is globular protein capable of inhibiting the function serine proteases, such as plasmin, trypsin, kallikrein and chymotrypsin [1–3]. Inhibition of the aforementioned proteins impedes the intrinsic pathway of fibrolysis and coagulation in humans which has led to extensive research and the development of a the protein based drug Trasylol [4–6]. This globular protein is characterized by a structure of two anti-parallel sheets and up to two helical regions being stabilized by three disulfide bridges [1-3]. The basic structure can be represented by the following simplified illustration (see Figure 1).

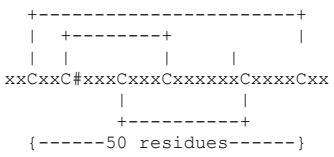

Figure 1 Simplified representation of the bovine pancreatic trypsin inhibitor domain. "C" denotes conserved cysteines that are involved in disulfide bonds. "#" denotes active site residue. Image: [8].

The arising of high throughput techniques dramatically accelerated the rate at which data are generated. To tackle these enormous data sets modern computational approaches are imperative. Machine learning approaches have been widely adopted in the bioinformatic context [9]. Conservation patterns of a given protein domain may be represented by Markov chains. They describe discrete-time stochastic processes and are the core of *hidden Markov models* (HMM). They are used in speech recognition as well as in the modelling of the likelihood of a residue to be present in a given position of a protein sequence. HMMs can be trained to identify a variety of patterns, which is necessary for domain profiling. Protein sequences of a certain family share the same function which is reflected in their structure and their evolutionary relationship. Some positions within a protein family are more conserved than others and are typically the ones crucial for a functional active site. These patterns can be exploited by a profile HMM. The topology of an HMM can be described as illustrated in Figure 2 [10,11]. The advantages of this approach are that it allows to characterize an entire protein family. Further it is based on formal statistics and probability and it allows to make libraries of hundreds of profile HMMs applicable to large scale data. Our aim was to make use of the HMM to create a method for identifying the bovine pancreatic trypsin inhibitor domain.

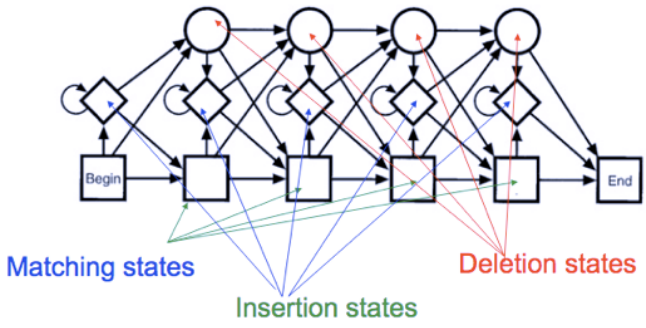

Figure 2 State diagram of an HMM for a sequence. The number of matching sates is given by the average sequence length in the family and is shown in blue. Insertion states are highlighted in green while deletion states are highlighted in red. Image modified from [11].

2 Methods

Data Retrieval for Training of the Profile Hidden Markov Model

Protein sequences containing the BPTI domain were downloaded from the *Protein Data Bank* (PDB, April 2020) [12,13]. We included only structures that met the following criteria: a resolution ranging between 0 – 3 Å, labeled with the Pfam (protein families) database identifier PF00014 associated to the BPTI domain, a polymer chain length of 50-70 residues [8,14]. We chose this particular range of resolution to ensure appropriate modeling of interacting bonds and the α-carbons. The query returned 39 protein structures.

BLASTclust was used for clustering sequences of a certain level of similarity to generate a non-redundant sequence set [15]. It takes as input sequences in FASTA format and returns a file containing one cluster per line. Clustering is achieved by computing all possible pairwise matches that are found by the BLAST algorithm [16–18]. The parameters were set for clustering sequences of 99% identity with a coverage of 0.99. From each cluster we selected the most representative structure according to the best resolution.

By scrutinizing the FASTA file generated from the multiple structural alignment in PDBe Fold v2.59. (src3) (April 2014) [19–21], we were able to identify erroneously classified sequences. These sequences were removed from the initial set. Then the clustering and the multiple structure alignment was repeated.

Training of the profile HMM

The profile HMM was trained on the basis of the sequence alignment derived from structural superposition. For generating the profile HMM we employed HMMER3.3 (Nov 2019) [22]. It is a program designed to identify distant homologs while depending on the strength of its underlying probability models. The program generates the profile with the help of a specific scoring system for deletions, substitutions and insertions based on the transition probability. To account for uncertainty HMMER3 calculates match scores by considering all possible alignments weighted by their relative likelihood [22,23]. We used this profile HMM to generate a sequence logo on the Skylign Web application illustrating the conserved residues (see Figure 3) [24,25].

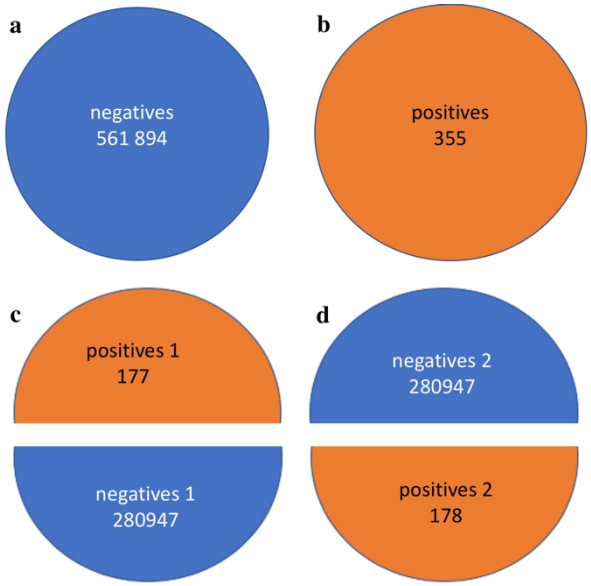

Figure 3 Illustration of negative and positive set. a and b, respectively. The sets were merged in the following fashion: (c)'positives 1' together with 'negatives 1'. (d) 'positives 2' were merged with 'negatives 2'. This way we could use one half for performance testing and the other half for optimization.

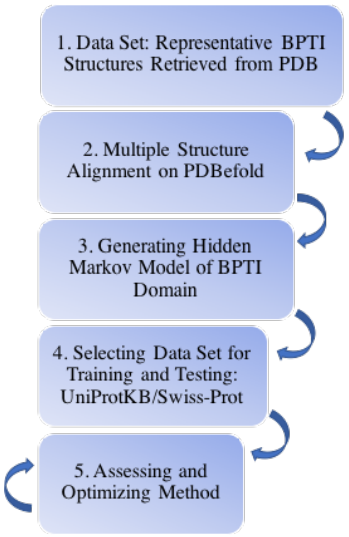

Figure 4 Schematized steps of the workflow used in this project.

Retrieving Data for Method Testing

For the testing of our profile HMM we had to use a dataset capable of validating the classification. The logic choice was using data from SwissProt, a part of UniprotKB containing only manually annotated proteins [26,27]. We retrieved protein sequences containing the BPTI domain for the positive set and proteins lacking such domain for the negative set. The query for the positive set was: "**Cross-references > Family and domain databases > Pfam (PF00014) AND reviewed:yes**" and returned 359 positive hits. The 561 894 negative hits were retrieved with the search term "**NOT Cross-references > Family and domain databases > Pfam (PF00014) Reviewed > Reviewed search term NOT database:(type:pfam pf00014) AND reviewed:yes**". Both sets were saved locally and reformatted as simplified FASTA.

It is important to note that for training and validation the positive set must exclude the sequences used in the training of the profile HMM. Given the search term, these sequences would however be part of the retrieved set. Further, sequences deposited in the UniprotKB often have multiple associated structures published in the PDB. Taking this into account we removed sequences sharing 100% identity from the positive set to avoid boosting the performance with a biased dataset. Running a position specific iterated BLAST via *blastpgp* enabled us to identify and remove them such that our final positive set contained 355 BPTI sequences.

Preparing the Data for 2-Fold Cross - Validation

To ensure our profile HMM is able to identify the BPTI domain, we tested it on a data set containing proteins with and without the domain.
We need to prepare and annotate the data available to be able to collect significant statistics for assessing the quality of our model/work.

After randomization, both the positive (355 sequences) and the negative set (561 894 sequences) were split. The negative set being an even number was split in equal halves of 280947 sequences while the positives had to be split in one smaller set containing 177 and the other containing 178 sequences. Ideally, we should be able to randomize them in two sets containing respectively one half of the negatives and one half of the positives.

For two-fold cross validation *hmmersearch* was employed [22]. *This* program takes as input an HMM file and searches this profile against a sequence database (a FASTA file). The profile HMM serves as the search term to browse the data base and find sequences akin to the family underlying the profile HMM [19,20]. We ran the *hmmsearch* tool with the options set for computing the E-value independent from the sample size enabling us to compare outcomes of different searches on our four sets. All the heuristic filters were turned off.

The tool was run four times, each time one of the sets of sequences described before was given as input, i.e.: 'positives 1', 'positives 2' 'negative 1' and then 'negative 2', see Figure 3 to retain information of the type (positive or negative) of the data. From the output of each run we extracted the fields containing the sequence ID of the targets and the E-value (of the best domain). We stored that information a file and added the information on kind of sequence (positive labelled with '1' or negative labelled with '0'). Since *hmmsearch* returned only data of sequences with E-values below 10 the ID's of sequences scoring higher than that were missing. We recovered those IDs and appended them to the output of *hmmsearch* and assigned a generic E-value of 10 which is high enough to classify them as negatives. All this was achieved via bash scripting.

Performance Testing by 2-Fold Cross - Validation

For testing our model, we merged the files containing the sequence ID, the E-value and the kind; into one positive half with another negative half of the set as illustrated in the Figure 3. We developed a python program that takes the aforementioned file and a threshold as an input and computes a confusion matrix which contains true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) (see Table 1). The program computed in addition the accuracy (ACC), Matthew Correlation (MCC), true positive rate (TPR), false positive rate (FPR) and the positive predictive value (PPV). Both a receiver operating characteristic (ROC) and a precision recall (PR) curve were plotted in pyplot.

		Actual class	
		Positives	Negatives
predicted class	Positives	TP	FP
	Negatives	FN	TN

Table 1 Confusion matrix showing the true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN).

(i) Accuracy

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

(ii) Matthew's Correlation Coefficient

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

(iii) Positive predictive value

$$PPV = \frac{TP}{TP + FP}$$

(iv) True positive rate

$$TPR = \frac{TP}{TP + FN}$$

Generating a Consensus sequence to Analyze False Positives and False Negatives

First a consensus sequence was generated using *hmmemit* with default parameters. It samples a sequence from a profile emitting a TP sequence for benchmarking or testing purposes. This sequence was saved to a file that also contained the FP and FN results. Then these sequences were aligned using *hmmalign* using the --trim option. The sequence alignment was visualized in Aliview (version 1.26) [28].

3 Results

The 25 sequences aggregated into 4 different multisequence clusters and the 10 remaining sequences were left by themselves forming singletons. One entire cluster of 2FJZ (Cluster **b** in grey see Table 2) had to be removed due to the lack of a BPTI domain. The four sequences had been wrongly annotated as BPTI but were actually harboring the Alzheimer's amyloid precursor (APP) domain instead. Re-clustering left us with the numeric black clusters as reported in Table 2. Blod font IDs were used for the final structure alignment.

Cluster	Blastclust output
1	**5PTI,** 1BPI, 4PTI, 6PTI, 9PTI
b	2FJZ, 2FK1, 2FK2, 2FMA
2	**1G6X**, 1K6U, 1QLQ
3	**1KTH,**1KNT, 2KNT
4-13	**6Q61, 3OFW, 5YV7, 1DTX, 1BPT, 1BTI, 1FAN, 1NAG, 7PTI, 8PTI**

*Table 2 Blastclust output shows which PDB ID's were clustered together. Cluster **b** (grey) was removed due to the lack of a BPTI domain. Clusters 5-14 are composed of single proteins. **Bold** font highlits the IDs that were used for the multiple sequence alignment in PDBefold.*

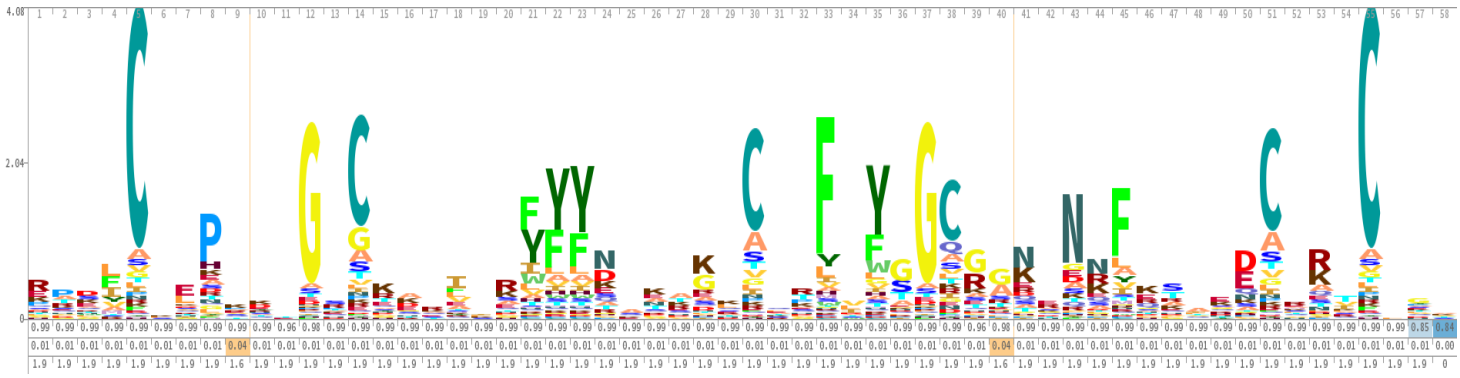

Figure 4 Sequence logo of the profile HMM of the BPTI domain. Image was generated using Skylign.org

Best Results Were Achieved at an E-value Between 10^{-8}-10^{-10}

The training returned the best outcome at an E-value threshold of 10^{-8}-10^{-10}. The results were identical for all three thresholds (Table 3). The entire Table of results is reported in the supplementary materials on github.

Set	Thr.	ACC	MCC	TPR	FPR	TNR	PPV	NPV
TS	10^{-8} - 10^{-10}	0.999	0.999	1.000	0.000	1.000	1.000	0.994
TS2	**10^{-8} - 10^{-10}**	**1.000**	**1.000**	**1.000**	**0.000**	**1.000**	**1.000**	**1.000**
VS	10^{-8} - 10^{-10}	0.999	0.994	0.988	0.000	1.000	1.000	0.999

Table 3 Statistics of testing (TS), corrected testing set (TS2) and validation set (VS). The three consecutive best performant E-value thresholds (Thr.) returned identical test statistics within each set. ACC: accuracy, MCC Matthews correlation coefficient, TPR: true positive rate, FPR: false positive rate, TNR: true negative rate, PPV: positive predictive value, NPV negative predictive value.

a) Training Set

Threshold 10^{-8}		Actual class: Positives	Actual class: Negatives
predicted class	Positives	TP = 177	FP = 1
	Negatives	FN = 0	TN = 280947

b) Validation Set

Threshold 10^{-8}		Actual class: Positives	Actual class: Negatives
predicted class	Positives	TP = 176	FP = 0
	Negatives	FN = 2	TN = 280947

Table 4 Confusion matrix of training set (a) and testing set (b). TP: true positives, FP: false positives, FN: false negatives, TN: true negatives.

False Positive in the Training Set is True Positive After All

The false positive obtained in the training set has the UniprotKB accession number G3LH89. The UniprotKB/Swissprot entry was titled Kunitz-type serine protease inhibitor Bi-KTI and had a heuristic annotation score of 3 (out of 5). It was not endowed with any Pfam identifier and Pfam was not crosslinked in the entry (June 2020) at all [29]. Investigating it closer we found that it was however annotated with the InterPro ID IPR002223 placing it into the pancreatic trypsin inhibitor Kunitz family. We performed an ID/accession number search in Pfam considering the possibility that the protein has merely not been crosslinked to Pfam within UniprotKB. This could be verified as a Pfam entry with the same UniprotKB accession number was found. We were able to verify that the entry has also been annotated with PF00014 (see in Pfam G3LH89) [30].

Corrected Training Set

Threshold 10^{-8}		Actual class: Positives	Actual class: Negatives
predicted class	Positives	TP = 178	FP = 0
	Negatives	FN = 0	TN = 280946

Table 5 Corrected training set. TP: true positives, FP: false positives, FN: false negatives, TN: true negatives.

The protein sequence was removed from the negative training set (280946) and placed into the positive set (178). The *hmmserarch* was repeated with the corrected data according to the new findings (see Table 3 above **TS2**).

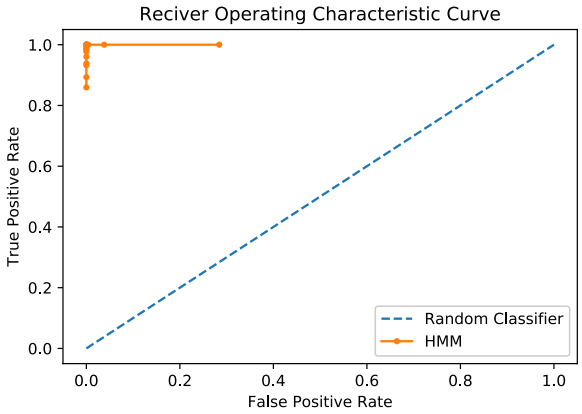

Figure 5 In orange the receiver operating characteristic curve obtained by our plotting our false positive rate (x - axis) against our true positive rate (y - axis). The blue dashed line corresponds to the hypothetical performance of a random classifier.

The receiver operating characteristic (ROC) curve was plotted (Figure 5). However, only 0.063% of our sequences in the data set contain the BPTI domain, making this dataset extremely imbalanced. Due to this imbalance the ROC curve obtained did not reach a FP rate of 1(no data in top right corner) and no TP rate of 0 (no data in bottom left corner.

In such cases a precision recall (PR) curve has been reported to be a better indicator of quality for a predictor [31]. The PR curve plotted can be seen in Figure 6.

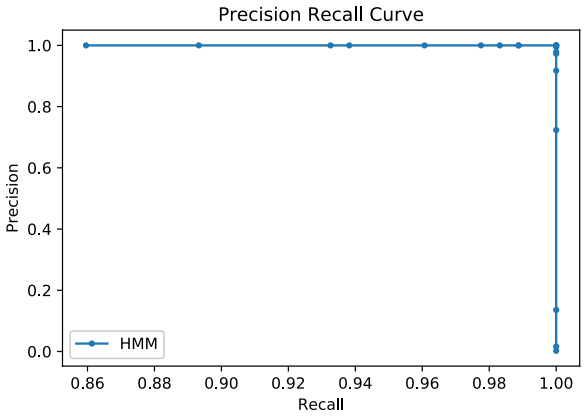

Figure 6 Plot of the precision recall curve of the binary classifyer we developed.

Analyzing False Negatives

The best results of the validation set still returned two false negatives. Their UniprotKB accession numbers are O62247 and D3GGZ. Both sequences belong to the phylum Nematoda and the Class Chromadorea. Our profile HMM was built on structures from 8 *Bos Taurus,* 1 *Homo sapiens, 2 Dendroaspis angusticeps*, 1 *Conus striatus* and one *Stichodactyla helianthus.* The wrongly classified proteins of the phylum Nematoda are at a distance of 5 clades from the last common ancestor Eumetazoa of the profile HMM. This overall distance and the sequences' distance to *Bos taurus* which contributed 61.5% of the BPTI sequences used for generating the model may explain the erroneous classification.

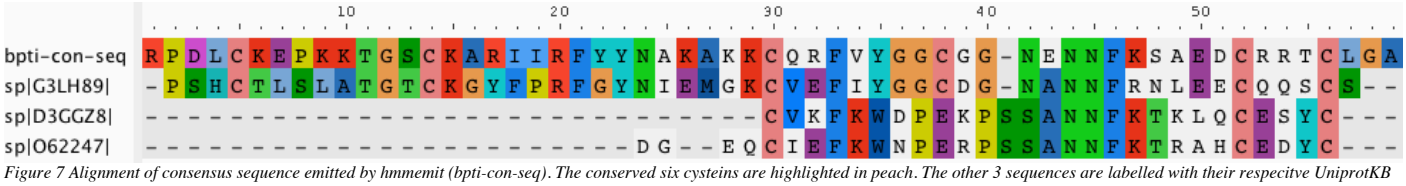

Figure 7 Alignment of consensus sequence emitted by hmmemit (bpti-con-seq). The conserved six cysteins are highlighted in peach. The other 3 sequences are labelled with their respecitve UniprotKB accession number. Image generated in Aliview version 1.26.

4 Conclusion

The aim of the project was to develop a profile Hidden Markov Model able to identify the bovine pancreatic trypsin inhibitor domain. We found that our method classifies the domains best at a threshold of 10^{-8}. While assessing the binary classification performance we found that the best threshold of 10^{-8} returned zero false positives and zero false negatives in our training set. This threshold did however return two false negatives which are evolutionary distant from the species that contributed their BPTI sequences to the training of the model. The performance is best illustrated in the graphical representation of the precision recall curve. Both precision and recall reached the maximum value of one. This is a favorable outcome considering the stark imbalance of positives (0.063%) and negatives (99.937%) in both training and testing sets.

Acknowledgements

Thanks to all professors who took their time to answer questions and to my colleagues who always engaged in stimulating discussions which have often led to new epiphanies.

Funding

This work has been solely supported by my savings acquired from previous fulltime employment as a diving instructor and later as a waitress.

Conflict of Interest: none declared.

References

1. Wlodawer A, Housset D, Kim KS, Fuchs J, Woodward C. Crystal structure of a Y35G mutant of bovine pancreatic trypsin inhibitor. J Mol Biol. 1991;220(3):757–70.

2. St Charles R, Padmanabhan K, Arni RV, Padmanabhan KP, Tulinsky A. Structure of tick anticoagulant peptide at 1.6 A resolution complexed with bovine pancreatic trypsin inhibitor. Protein Sci. 2000 Feb;9(2):265–72.

3. Biemann K, Papayannopoulos IA. Amino acid sequence of a protease inhibitor isolated from Sarcophaga bullata determined by mass spectrometry. Protein Sci. 1992;1(2):278–88.

4. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-672.

5. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research. 2018 Jan 4;46(D1):D1074–82.

6. Aprotinin [Internet]. [cited 2020 May 17]. Available from: https://www.drugbank.ca/drugs/DB06692

7. Nixon A, Wood CR. Engineered protein inhibitors of proteases. Current Opinion in Drug Discovery & Development. 2006;9(2):261–8.

8. Summary: Kunitz/Bovine pancreatic trypsin inhibitor domain [Internet]. [cited 2020 May 10]. Available from: http://pfam.xfam.org/family/PF00014

9. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min [Internet]. 2017 Dec 8 [cited 2020 May 11];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660/

10. Momand J, McCurdy A. Concepts in bioinformatics and genomics. International ed. New York: Oxford University Press; 2017. xvi+448.

11. HIDDEN MARKOV MODELS IN MULTIPLE ALIGNMENT. 2 HMM Architecture Markov Chains What is a Hidden Markov Model(HMM)? Components of HMM Problems of HMMs. - ppt download [Internet]. [cited 2020 May 23]. Available from: https://slideplayer.com/slide/4970773/

12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

13. Bank RPD. RCSB PDB [Internet]. [cited 2020 Jun 12]. Available from: https://www.rcsb.org/

14. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.

15. National Center for Biotechnology Information (NCBI) Documentation of the BLASTCLUST-algorithm. BlastClust [Internet]. [cited 2020 Jun 14]. Available from: https://ftp.ncbi.nih.gov/blast/documents/blastclust.html

16. Download BLAST Software and Databases Documentation [Internet]. [cited 2020 May 19]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

17. National Center for Biotechnology Information (NCBI): Documentation of the BLASTCLUST-algorithm. [Internet]. [cited 2020 May 20]. Available from: https://ftp.ncbi.nih.gov/blast/documents/blastclust.html

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct 5;215(3):403–10.

19. Krissinel E, Henrick K. Multiple Alignment of Protein Structures in Three Dimensions. In: R. Berthold M, Glen RC, Diederichs K, Kohlbacher O, Fischer I, editors. Computational Life Sciences [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005 [cited 2020 May 20]. p. 67–78. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 3695). Available from: http://link.springer.com/10.1007/11560500_7

20. PDBefold.pdf [Internet]. [cited 2020 May 20]. Available from: https://www.ebi.ac.uk/pdbe/docs/Tutorials/workshop_tutorials/PDBefold.pdf

21. PDBe < Fold < EMBL-EBI [Internet]. [cited 2020 May 20]. Available from: https://www.ebi.ac.uk/msd-srv/ssm/

22. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013 Jul 1;41(12):e121–e121.

23. HMMER [Internet]. [cited 2020 May 20]. Available from: http://hmmer.org/

24. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics. 2014 Jan 13;15(1):7.

25. Skylign [Internet]. [cited 2020 May 19]. Available from: http://skylign.org/

26. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

27. UniProt [Internet]. [cited 2020 Jan 6]. Available from: https://www.uniprot.org/

28. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014 Nov 15;30(22):3276–8.

29. Kunitz-type serine protease inhibitor Bi-KTI G3LH89, Bombus ignitus (Bumblebee) [Internet]. [cited 2020 Jun 12]. Available from: https://www.uniprot.org/uniprot/G3LH89

30. Pfam: Protein: VKT_BOMIG (G3LH89) [Internet]. [cited 2020 Jun 13]. Available from: http://pfam.xfam.org/protein/G3LH89

31. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One [Internet]. 2015 Mar 4 [cited 2020 Jun 1];10(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/