Bioinformatics, 2025, 1–7
doi: 10.1093/bioinformatics/052809
Advance Access Publication Date: 19th May 2025
Manuscript Category

Report

Modeling the BPTI/Kunitz Domain: A Profile HMM Approach for Protein Domain Annotation

Priscilla Castro-Vargas[1,*]

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum – Università di Bologna, Via F. Selmi 3 - 4th Floor, Room 183.

[1]https://orcid.org/0000-0001-6333-0419

*To whom correspondence should be addressed.

Associate Editor: Emidio Capriotti

Received on 19-05-2025; revised on 19-05-2025; accepted on

Abstract

Motivation: The BPTI/Kunitz domain, a cysteine-rich motif involved in protease inhibition and diverse biological processes, is of significant interest for drug discovery and functional annotation. This project addresses the need for a sensitive and specific detection method based on structural conservation.
Results: A profile Hidden Markov Model (HMM) was developed from a multiple structural alignment of curated Kunitz domain structures. The model demonstrated high predictive performance, achieving Matthews Correlation Coefficients (MCCs) up to 0.99 across two validation datasets. These results underscore the effectiveness of structure-informed HMMs as accurate tools for domain identification, with implications for functional annotation and computational biology pipelines.

Availability: All datasets, alignment files, and evaluation scripts are available at GitHub Repository.

Contact: priscilla.castro@studio.unibo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Kunitz-type protease inhibitor family is named after the bovine pancreatic trypsin inhibitor (BPTI) that was crystallized and isolated by Kunitz and Northrop in 1935 (1,2). The Kunitz-type domains are widely found in nature, they are cysteine-rich peptide chain composed of 50 to 70 amino acids and share a conserved structural fold that includes two antiparallel β-sheets and one or two helical segments, stabilized by three disulfide bonds (**Fig. 1**) (3,4).

Kunitz-type domains act as serine protease inhibitors through a standard inhibition mechanism that involves the formation of a tight, non-covalent interaction with the target enzyme (5). These inhibitors bind directly to the active site of the serine protease active site without inducing conformational changes, he interaction typically results in the formation of an antiparallel β sheet between enzyme and inhibitor (5,6). The specificity of inhibition is largely determined by the amino acid residues present at the reactive sites of the Kunitz domain (5).

In the biological context, Kunitz domains are versatile and have evolved to perform a variety of functions, many of which involve the inhibition of protease activity and the modulation of biological processes. Some Kunitz domains are involved in immunomodulation (7), anticoagulation (8) and defense against pathogens, among other functions (9). These properties have led to their potential use in developing drugs for treating several diseases and conditions (10). Accurate identification of such domains is essential for functional annotation, drug design and evolutionary studies.

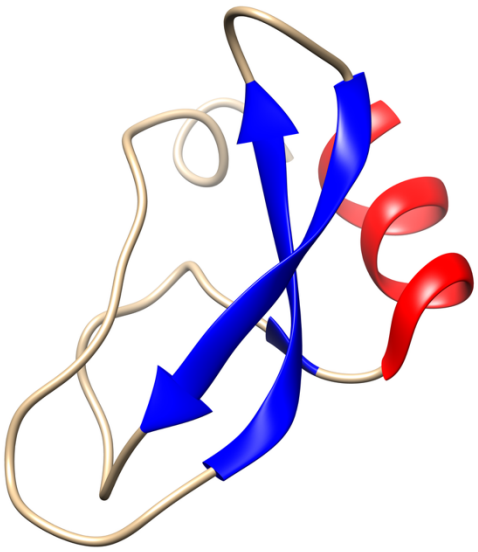

Figure 1. Kunitz domain. Representation of the 56 amino acids that form the Kunitz domain present in the protease inhibitor domain of Alzheimer's amyloid β-protein precursor (PDB entry – 1AAP) (11) made with UCSF Chimera (version 1.20) (12). The antiparallel β-sheets are presented in blue and the α-helix in red.

Hidden Markov Models (HMMs) are statistical models that capture hidden information from observable sequential symbols (13,14). Profile Hidden Markov Models (HMMs) are well-established tools for modeling protein families and domains due to their ability to capture position-specific residue conservation and variability (15). An effective HMM accurately represents the underlying biological process that generates the observed data and can also simulate that source realistically (14). To cite a few, HMM have been used for detecting β-structural motifs in proteins (16), and for predicting and identifying Src Homology 2 (SH2) domains within the proteome (17).

This project focuses on the development of a computational method based on a profile HMM derived from structurally aligned protein domains to detect the presence of the BPTI/Kunitz domain in protein sequences. By leveraging structural information, the model aims to annotate the Kunitz domain with high sensitivity and specificity.

2 Materials and Methods

A simplified diagram of the workflow can be found in **Fig. 2**.

2.1 Selection of Protein Sequences and Structures

The protein sequences were retrieved from UniProtKB/Swiss-Prot (18) on April 1st, 2025 using the following filters: 1. Taxonomy: Homo sapiens (NCBI Taxonomy ID: 9606), 2. Reviewed: Yes, 3. Pfam Domain: PF00014 (Kunitz-type protease inhibitor). The results were three datasets: human proteins with the Kunitz domain (HK), human proteins without the Kunitz domain (HNK) and proteins with the Kunitz domain that were not human (NHK). Additionally, all the protein sequences of the UniProtKB/Swiss-Prot (18) were previously retrieved on March 18th, 2025. These datasets were used to construct positive and negative test sets for model validation, ensuring that training sequences were excluded from the validation set.

On the other hand, structures of Kunitz domains were obtained from the Protein Data Bank (PDB) (19) on April 7th, 2025 using the following criteria: 1. Pfam Annotation: PF00014, 2. Data Collection Resolution ≤ 3.5 Å and Polymer Entity Sequence Length 45-80 amino acids. The structures were clustered using CD-HIT (version 4.8.1) (20), with the default behavior (90% sequence identity threshold) to achieve a non-redundant set of protein sequences and select representative sequences for downstream analysis.

2.2 Multiple Sequence Alignment and Multiple Structural Alignment

A multiple structural alignment of the clustered sequences was generated using PDBeFold (21) to validate structural similarities and identify potential outliers. The results of this alignment were visualized with UCSF Chimera (version 1.20) (12). The resulting alignment was inspected to evaluate number of aligned residues, the root-mean-square deviation (RMSD), used to measure the difference between structures (22), the Secondary Structure Elements (SSEs) and the Q-score, that estimates the quality of each match. The final multiple sequence alignment was exported in FASTA format for HMM generation.

2.3 Preparation of the Datasets

A single FASTA file with all the sequences from HK and NHK was made (AK), for a total of 397 positive sequences. Using BLAST's (version 2.16.0+) (23) makeblastdb a BLAST database was created with the AK sequences. With blastp, the sequences of the proteins used for the multiple sequence alignment were searched against the AK database. All sequences within defined thresholds (sequence identity ≥ 95% and number of aligned residues ≥ 50) were removed from the positive dataset to avoid biases in the validation, for a final amount of 368 positive sequences.

In addition, all the Kunitz sequences were extracted from the file with all the protein sequences of Uniprot/Swiss-Prot to conform the negative dataset, for a total amount of 572,573 negative sequences. Each of these datasets was randomized and partitioned in two datasets, each one

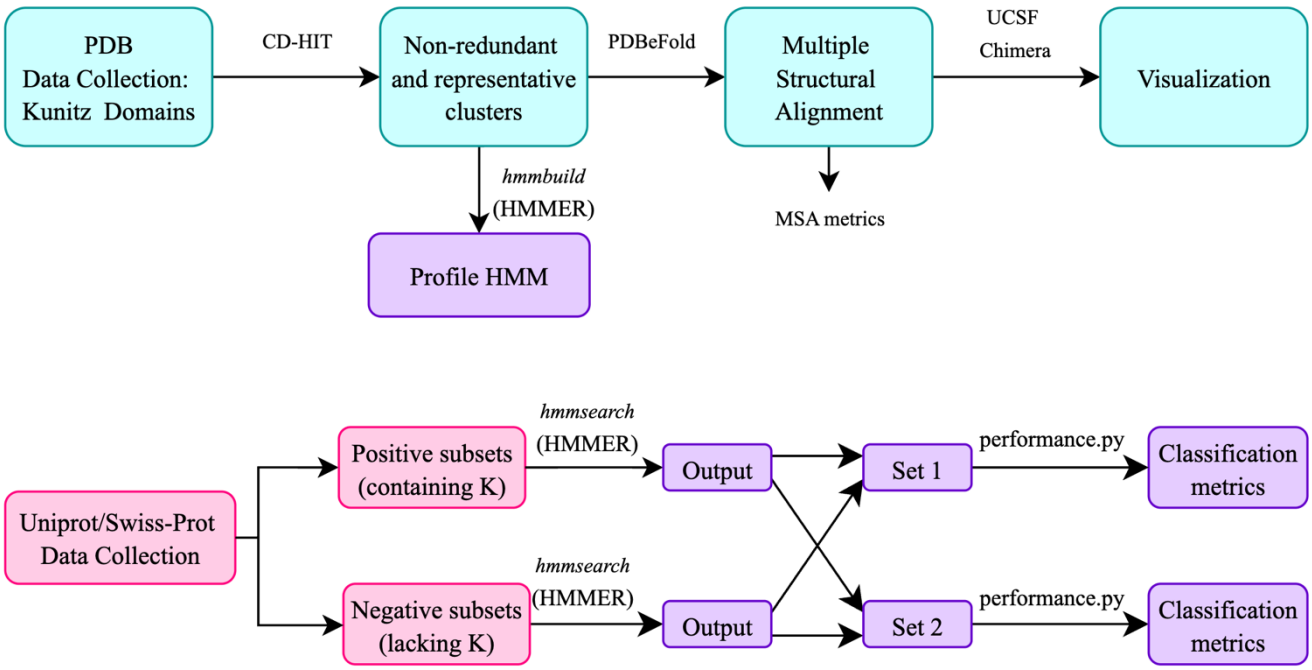

Figure 2. Simplified workflow of the investigation. The data collection was made from the PDB and the Uniprot/Swiss-Prot websites. The structures were clustered and aligned. The alignment was then visualized. The clusters were used to build a profile HMM that was validated with sets obtained from Uniprot/Swiss-Prot. The validation results were analyzed, and classification metrics were obtained (Diagram made with https://app.diagrams.net).

containing half of the original dataset. For a total amount of four subsets, two positives (pos_1 and pos_2) and two negatives (neg_1 and neg_2). The K-fold cross-validation test optimizes the classification threshold of the model (15), in this research the 2-fold cross validation was used.

2.4 Construction, Validation and Evaluation of the Profile Hidden Markov Model

The profile HMM was built using HMMER's (version 3.4) (24) hmmbuild from the cleaned and non-redundant multiple structural alignment. This profile was evaluated using hmmsearch to search the sequences of the four datasets (two positives and two negatives) against the profile HMM. The tabular summary output of this searches was saved. From these results, two sets were formed, each containing one of the positive outputs and one of the negative outputs. These sets contained sequence ID, class label (0 for negative and 1 for positive), bit-score (a measure of how well a sequence matches the profile HMM (25)), and E-value (statistical significance of the bit-score (25)).

To evaluate the classification performance of the profile HMM a python (version 3.13) script was developed (available from the corresponding author). This script builds a confusion matrix, a table built with true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) (26). The script also calculates common classification performances, such as: quartile-based accuracy (Q2), sensitivity (True Positive Rate), precision (Positive Predictive Value), and Matthew's Correlation Coefficient (MCC) (27).

The MCC is a balanced metric that maximizes all four basic rates (TP, TN, FP, FN), with respect to the scale, the worst and minimum value is -1 and the best and maximum values is +1. Recently, a study explaining why the MCC should replace the ROC AUC as standard in scientific studies involving a binary classification (28).

3 Results and Discussion

3.1 Data Collection of Protein Sequences and Structures

The Uniprot search yielded 18 human proteins containing the Kunitz domain, 20,400 human proteins lacking the Kunitz domain and 379 non-human proteins containing the Kunitz domain.

Additionally, the PDB search returned 158 structures, that were clustered into 25 representative sequences. Three structures were excluded from the downstream analysis: one due to its length of 127 residues, one given its length of 78 residues, it is probable that the Kunitz domain its connected to a larger chain; and the other for having only 38 residues, it's unlikely that the last one even corresponds to Kunitz domains. For a total amount of 22 non redundant and representative structures. In this investigation, the

assumption that the Pfam ID was correct was made, but the possibility of mislabeling exists and could affect the subsequent generation of the profile HMM. Alternatives to retrieve the structures include using a prototype structure to search for similar ones in the PDB, finding a protein in UniProt that has an annotated BPTI/Kunitz-type domain and has a corresponding 3D structure, attempting a direct search in the PDB for proteins with structurally resolved Kunitz domains.

Several consideration must be taken into account when selecting domains. For example, PDB files may include multiple chains, each chain can feature one or more domains, which may be of the same or different types, or the same protein might appear in different PDB entries. Taking into account these reasons, for the profile HMM model only domains that were on a single chain of each PDB entry were considered.

3.2 Multiple Sequence Alignment and Multiple Structural Alignment

The result of the multiple structural alignment can be visualized in **Fig. 3** and the multiple structural alignment can be seen **Fig. 4**. The complete multiple alignment results can be retrieved from the following GitHub Repository. The overall number of aligned residues were 50, which corresponds with the known dimensions of the Kunitz domain (10). The overall number of aligned SSEs was 3, which corresponds to the antiparallel β-sheets and the α-helix (as shown in **Fig. 3**). Additionally, the overall RSMD was 1.019 Å, which means that the 3D conformations of the aligned parts are nearly identical. Finally, the Q-score was 0.6289, which indicates moderate to good quality. There are some regions with ambiguities or misalignment, possibly the ones between the stablished secondary structures.

Figure 3. 3D Multiple Structural Alignment. Representation of the multiple structural alignment of the 22 non redundant and representative structures of the Kunitz domain made with UCSF Chimera (version 1.20) (12).

As illustrated in **Fig. 4**, there level of residue conservation across the alignment is variable, leading to differing degrees of consensus in each position. A noteworthy feature of the results is the high conservation of cysteine residues. Considering that Kunitz domains are structurally stabilized by disulfide bonds (4,10), this finding supports the notion that the disulfide bonds formed by these cysteine residues are evolutionary conserved.

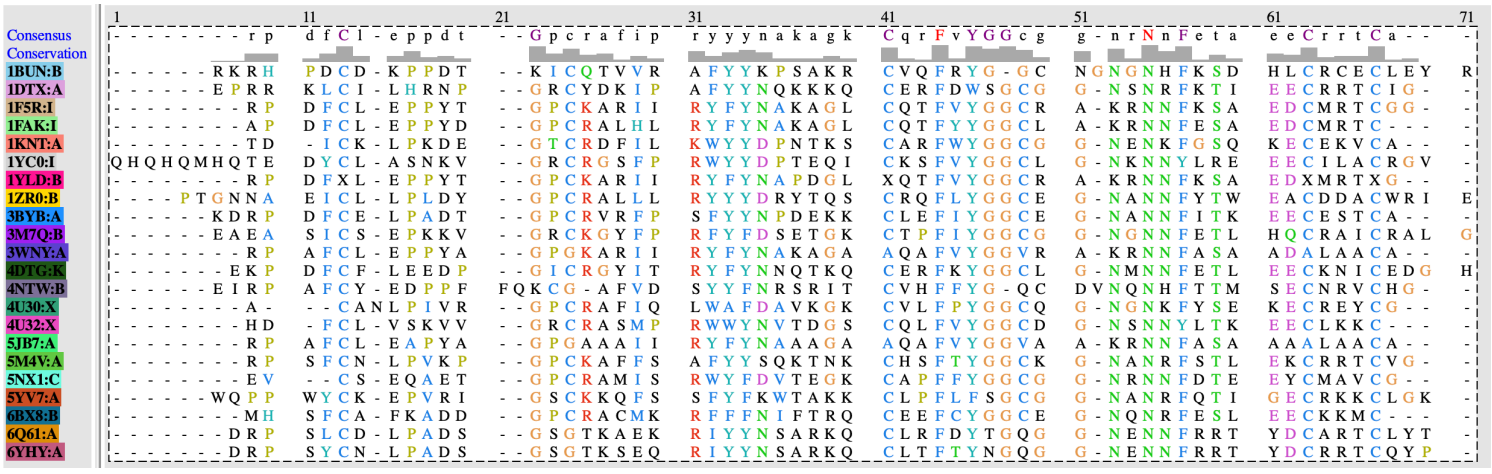

Figure 4. Multiple Sequence Alignment. Representation of the multiple sequence alignment of the 22 non redundant and representative structures of the Kunitz domain made with PDBeFold and represented with UCSF Chimera (version 1.20). The color of each PDB identifier corresponds to the colors on the 3D Representation of the Multiple Structural Alignment.

3.3 Evaluation of the profile HMM

The evaluation of the performance was made testing different thresholds (1e-01 to 1e-12) and producing classification matrices. As seen in **Fig. 5**, the MCC values ranged approximately from 0,92 to 0,99. Both sets exhibited a similar trend corresponding to an initial increase in MCC as the threshold becomes more stringent (from e-value in the range 1e-01 to 1e-5), followed by a plateau or slight decline. Specifically, both sets reached its maximum MCC around an e-value of 1e-5-1e-6. This behavior suggests that a more stringent e-value threshold improves classification performance up to a point, after which the benefit plateaus or slightly diminishes. The high MCC values across the range indicate consistently strong predictive performance both sets.

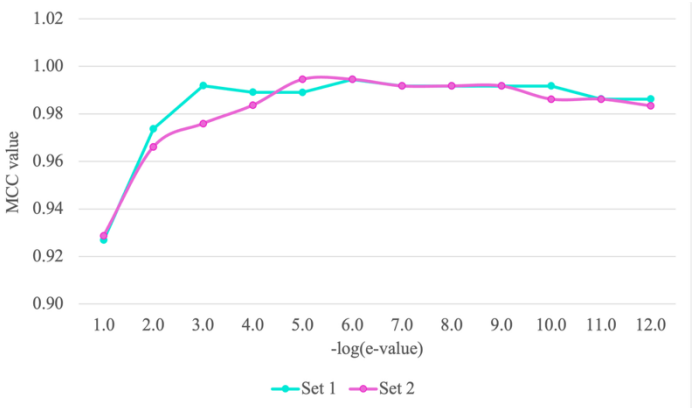

Figure 5. Variation of the MCC value as a function of the threshold. The graph illustrates the behavior of the MCC value of the Set 1 (cyan) and Set 2 (fuchsia) for different e-values.

The classification performance of the profile HMM applied to set 2 using a e-value threshold of 1e-05 is presented in **Table I**. These results suggest strong overall classification performance, with near-perfect accuracy and MCC, consistent with the high MCC values observed in **Fig. 5**.

Table I. Confusion matrix results for set 2 using an e-value of 1e-05.

		Gold standard	
		Positive Kunitz sequence	Negative Kunitz sequence
Test outcome	Positive Kunitz sequence	180	0
	Negative Kunitz sequence	2	286,416

Some known Kunitz domains were missed by the model, possibly due to sequence divergence or structural variation not captured in the seed alignment. These false negatives highlight the model's sensitivity to the choice of training sequences and suggest that increasing the diversity of the alignment may enhance recall.

Overall, the results indicate that a structure-informed HMM can identify Kunitz domains with high accuracy, but care must be taken to interpret borderline cases and to assess database completeness.

4 Conclusion

This study demonstrates the efficacy of a structure-informed profile Hidden Markov Model for the identification of the BPTI/Kunitz domain in protein sequences. By leveraging multiple structural alignments of high-quality, non-redundant domain instances, the model achieved high sensitivity and specificity, as reflected by Matthews Correlation Coefficient (MCC) values approaching 0.99. The conserved structural features—particularly the disulfide-stabilizing cysteine residues—were effectively captured by the alignment, reinforcing the domain's evolutionary conservation. The robust performance of the model across diverse validation datasets highlights the power of combining structural and sequence-based information for protein domain annotation. While some limitations persist, particularly in detecting divergent sequences, the approach provides a strong foundation for further applications in proteome-wide domain identification and functional annotation pipelines.

Acknowledgements

Special thanks to Dr. Capriotti for his invaluable guidance throughout this research.

Funding

This work has been supported by Facultdad de Microbiología (Universidad de Costa Rica) and Dipartamento di Farmacia e Biotecnologie (Alma Mater Studiorum – Università di Bologna).

Conflict of Interest: none declared.

References

1. Kunitz M, Northrop JH. ISOLATION FROM BEEF PANCREAS OF CRYSTALLINE TRYPSINOGEN, TRYPSIN, A TRYPSIN INHIBITOR, AND AN INHIBITOR-TRYPSIN COMPOUND. J Gen Physiol. 1936 Jul 20;19(6):991–1007.

2. M Laskowski Jr, Kato I. Protein Inhibitors of Proteinases. Annu Rev Biochem. 1980 Jul 1;49(Volume 49, 1980):593–626.

3. Buchanan A, Revell JD. Chapter 8 - Novel Therapeutic Proteins and Peptides. In: Singh M, Salnikova M, editors. Novel Approaches and Strategies for Biologics, Vaccines and Cancer Therapies [Internet]. San Diego: Academic Press; 2015 [cited 2025 May 17]. p. 171–97. Available from: https://www.sciencedirect.com/science/article/pii/B9780124166035000080

4. Mishra M. Evolutionary Aspects of the Structural Convergence and Functional Diversification of Kunitz-Domain Inhibitors. J Mol Evol. 2020 Sep 1;88(7):537–48.

5. Ranasinghe S, McManus DP. Structure and function of invertebrate Kunitz serine protease inhibitors. Dev Comp Immunol. 2013 Mar 1;39(3):219–27.

6. Morais KLP, Ciccone L, Stura E, Alvarez-Flores MP, Mourier G, Driessche MV, et al. Structural and functional properties of the Kunitz-type and C-terminal domains of Amblyomin-X supporting its antitumor activity. Front Mol Biosci [Internet]. 2023 Feb 9 [cited 2025 May 18];10. Available from: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2023.1072751/full

7. de Magalhães MTQ, Mambelli FS, Santos BPO, Morais SB, Oliveira SC. Serine protease inhibitors containing a Kunitz domain: their role in modulation of host inflammatory responses and parasite survival. Microbes Infect. 2018 Oct 1;20(9):606–9.

8. Sun F, Deng X, Gao H, Ding L, Zhu W, Luo H, et al. Characterization of Kunitz-Domain Anticoagulation Peptides Derived from Acinetobacter baumannii Exotoxin Protein F6W77. Toxins. 2024 Oct;16(10):450.

9. Arnaiz A, Talavera-Mateo L, Gonzalez-Melendi P, Martinez M, Diaz I, Santamaria ME. Arabidopsis Kunitz Trypsin Inhibitors in Defense Against Spider Mites. Front Plant Sci [Internet]. 2018 Jul 10 [cited 2025 May 18];9. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2018.00986/full

10. Buchanan A, Revell JD. Chapter 8 - Novel Therapeutic Proteins and Peptides. In: Singh M, Salnikova M, editors. Novel Approaches and Strategies for Biologics, Vaccines and Cancer Therapies [Internet]. San Diego: Academic Press; 2015 [cited 2025 May 18]. p. 171–97. Available from: https://www.sciencedirect.com/science/article/pii/B9780124166035000080

11. Hynes TR, Randal M, Kennedy LA, Eigenbrot C, Kossiakoff AA. X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid .beta.-protein precursor. Biochemistry. 1990 Oct 1;29(43):10018–22.

12. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605–12.

13. Eddy SR. What is a hidden Markov model? Nat Biotechnol. 2004 Oct;22(10):1315–6.

14. Franzese M, Iuliano A. Hidden Markov Models. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology [Internet]. Oxford: Academic Press; 2019 [cited 2025 May 18]. p. 753–62. Available from: https://www.sciencedirect.com/science/article/pii/B9780128096338204883

15. Srivastava PK, Desai DK, Nandi S, Lynn AM. HMM-ModE – Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinformatics. 2007 Mar 27;8:104.

16. Kumar A, Cowen L. Recognition of beta-structural motifs using hidden Markov models trained with simulated evolution. Bioinformatics. 2010 Jun 15;26(12):i287–93.

17. Jablonowski K. Hidden Markov Models for Protein Domain Homology Identification and Analysis. Methods Mol Biol Clifton NJ. 2017;1555:47–58.

18. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025 Jan 6;53(D1):D609–17.

19. Burley SK, Bhatt R, Bhikadiya C, Bi C, Biester A, Biswas P, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. Nucleic Acids Res. 2025 Jan 6;53(D1):D564–74.

20. CD-HIT: accelerated for clustering the next-generation sequencing data | Bioinformatics | Oxford Academic [Internet]. [cited 2025 May 17]. Available from: https://academic.oup.com/bioinformatics/article/28/23/3150/192160

21. 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. Berlin, Heidelberg: Springer; 2005. p. 67–78.

22. Wang X, Snoeyink J. Multiple structure alignment by optimal RMSD implies that the average structure is a consensus. Comput Syst Bioinforma Comput Syst Bioinforma Conf. 2006;79–87.

23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

24. HMMER web server: 2018 update | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2025 May 17]. Available from: https://academic.oup.com/nar/article/46/W1/W200/5037715?login=false

25. EMBL-EBI. Threshold adjustment | Pfam [Internet]. [cited 2025 May 19]. Available from: https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/modelling-in-pfam/threshold-adjustment/

26. de Giorgio A, Cola G, Wang L. Systematic review of class imbalance problems in manufacturing. J Manuf Syst. 2023 Dec 1;71:620–44.

27. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000 May 1;16(5):412–24.

28. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. 2023 Feb 17;16(1):4.