Development of a Profile Hidden Markov Model for Kunitz-Type Protease Inhibitor Domain Detection

Cagnini Luca[1]

[1]Department of Pharmacy and Biotechnology, University of Bologna, Italy.

Bioinformatics Master's Degree Course

Abstract

The Kunitz domain is a well-known protease inhibitor motif. It is found ubiquitously in proteins from mammals to invertebrates. Its small size, structural conservation and biological relevance make it a key target in protein annotation and therapeutic development.

This study presents a developed structure-informed Hidden Markov Model (HMM) to identify Kunitz domains in protein sequences. The profile HMM was built using high-resolution structural data from PDB and sequence filtering strategies (using BLAST and CD-HIT). Upon training, the HMM functioned as a binary classifier, facilitating precise annotation of the presence or absence of the Kunitz domain within protein sequences obtained from UniProt/SwissProt. The model was evaluated on curated positive and negative datasets, achieving an important result in classification performance (MCC), confirming its reliability for detecting Kunitz domains.

Contact: luca.cagnini@studio.unibo.it

Supplementary information: Supplementary materials are available at: https://github.com/LucaCagnini/HMM-Kunitz

1. Introduction

1.1 the Kunitz Domain

The Kunitz-type domain (Pfam: PF00014, InterPro: IPR002223) is a compact protein motif of ~60 amino acids that adopts a conserved α+β fold, stabilized by three disulfide bonds with a characteristic C1–C6, C2–C4, and C3–C5 pattern (Fratini et al. 2022). These disulfide bridges play a dual role in stabilizing the native conformation of the domain and facilitating its interaction with proteases, as they are involved in forming the protease-binding loop. For this reason, it is a reliable scaffold for protease inhibition and other biological functions.

Despite their conserved structure, Kunitz protein domains are found in diverse organisms serving a variety of roles: form inhibition of serine proteases to the modulation of ion channels in venomous animals, to immune evasion in parasites (H. Zhang et al, 2021). Their conservation and functional relevance highlight the interest in the development of computational tools for their detection. By targeting proteases implicated in pathological conditions, such as cancer, inflammation, and neurodegenerative disorders, Kunitz protein inhibitors hold the potential to develop targeted therapies with improved efficacy and reduced side effects (S Ranasinghe and DP McManus, 2013).

Kunitz domains are stable peptides able to recognize specific protein structures, working as competitive protease inhibitors in their free form. These properties have led to attempts at developing biopharmaceutical drugs from Kunitz domains.

1.2 Hidden Markov Models (HMMs)

Hidden Markov Models are statistical based methods widely used in biology to model position-specific conservation and variability profiles in aminoacidic and nucleotides sequences, also widely used for domain annotation (a.g., PFAM) (K. Jablonowski., 2017). In this study we developed a Profile HMMs built from aligned sequences, to capture the signature of the Kunitz domain. While a sequence profile represents the consensus of aligned sequences, a profile-HMM enhances this by incorporating

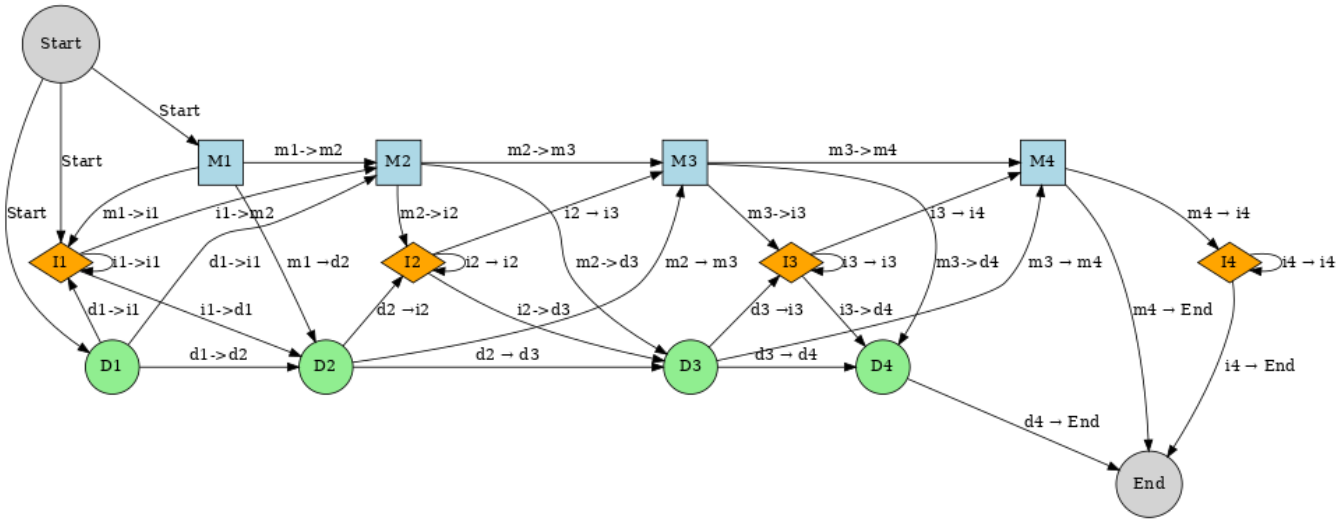

Fig. 1. Graph representation of a Hidden Markov Model architecture of four states and three layers: M (1,4) matches; I (1,4) insertions; D (1,4) deletions. Arrows indicates possible transitions between nodes, with possible transition and emission probabilities associated with them.

probabilistic modelling, to account for evolutionary variability (B.J. Yoon., 2009). The challenges in accurately predicting the presence of the Kunitz domain within protein sequences arise due to the variations in the domain's structure and function across different species (Eugene Krissinel and Kim Henrick., 2004).

An HMM (figure 1) has the following components:

- States: in protein sequences, a state might represent a particular residue position in a conserved domain.
- Transition Probabilities: The probability of moving from one state to another.
- Emission Probabilities: The probability of emitting a particular observable symbol from a hidden state.
- Initial Probabilities: The probability distribution over which state the model starts in.

HMMER is a standard suite used for building and searching HMM profiles (Martin Larralde et al.,2022).

1.3 Objective

The goal of this project is to develop a classical bioinformatics pipeline to construct and evaluate an HMM-based method for detecting the Kunitz domain. The model is trained on structurally aligned sequences, tested on filtered datasets, and finally evaluated using statistical performance metrics using state of the art methods robust and tested in literature (Travis J Wheeler, Jody Clements, 2014).

2. Methods

2.1 Dataset Collection

Protein structures containing the Kunitz domain were retrieved from UniProt (https://www.uniprot.org/) using the following query:

Data Collection Resolution <= 3.5 AND (Identifier = "PF00014" AND Annotation Type = "Pfam") AND Polymer Entity Sequence Length <= 80 AND Polymer Entity Sequence Length >= 45

From the advance search 158 sequences were found. Results were saved into a .csv file containing protein identifier, Structure Data and Polymer Entity Data. The retrieved data was processed to extract amino acid sequences and reduce redundancy using CD-HIT (Cluster Database at High Identity with Tolerance) at a 90% identity threshold, yielding a representative non-redundant dataset. After the process, 25 clusters were obtained. Before proceeding, the clustered file was manually analyzed and filtered eliminating clusters with longer sequences. A representative sequence for each cluster was extracted. Results were saved on a. fasta file.

2.2 Multiple Sequence Alignment (MSA)

Structure-based multiple sequence alignment was performed using PDBeFold (https://www.ebi.ac.uk/msd-srv/ssm/) on the filtered sequences. The alignment was downloaded in. ali format, converted to FASTA, and formatted for compatibility with HMMER using custom awk scripts (supplement material). The results of the MSA were filtered considering RSMD (Root Squared Median Deviation) values lower than 1.

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|^2} \qquad (1)$$

RSMD is the most used quantitative measure of the similarity between two superimposed atomic coordinates. Of the 24 alignments, we selected 23 representative structures. Q-score was also considered: the Q-score is a measure used in structural alignment to assess how well two protein structures match in 3D space.

2.3 HMM construction

The profile HMM was built using hmmbuild (HMMER v3), which estimates model parameters from the MSA. The model captures the conserved features of the Kunitz domain, such as the disulfide-bonded cysteines and other key residues. Results were saved as .hmm file.

2.4 Dataset Preparation

To evaluate the performance of our model we started creating two data sets:
Positive set: non-redundant Kunitz-containing sequences from UniProtKB/SwissProt, excluding sequences used in the HMM training or with ≥95% identity to them (identified using blastp).
Negative set: SwissProt proteins not annotated with PF00014
Both datasets were shuffled and split into two subsets for 2-fold cross-validation. 2-fold cross-validation is used to ensure that the model's performance is robust, generalizable, and not overfitted with any subset of data.

2.5 HMM search

The hmmsearch command was used to evaluate all positive and negative sets using the trained model. Outputs were saved in .out and .class formats, recording sequence ID, class label, bit score, and E-value. Positive and negative sets were united to create two sets for the evaluation.

2.6 Performance Evaluation

Classification performance was assessed using a custom *performance.py* script, evaluating specific metrics.
Accuracy: the proportion of total predictions (both positive and negative) that are correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

Sensitivity (Recall): the proportion of actual positives that are correctly identified.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad (3)$$

Specificity: the proportion of actual negatives that are correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad (4)$$

Precision (PPV): the proportion of predicted positives that are positive.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (5)$$

Matthews Correlation Coefficient (MCC): A balanced score that considers all four values: TP, TN, FP, FN.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (6)$$

Performance was tested at various E-value (from 1e-1 to 1e-30), and the best threshold was selected based on MCC.

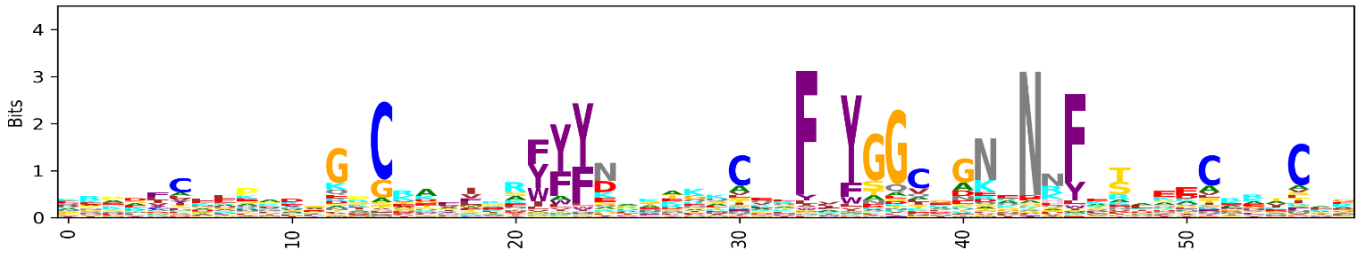

Fig. 2. Sequence logo representing the conserved residues within the Kunitz domain across multiple protein sequences. The height of each letter indicates the relative frequency of the corresponding amino acid at that position. Highly conserved cysteine residues, essential for forming disulfide bridges, are prominently displayed at positions 6, 15, 31, 39, 52, and 55 in the multiple sequence alignment.

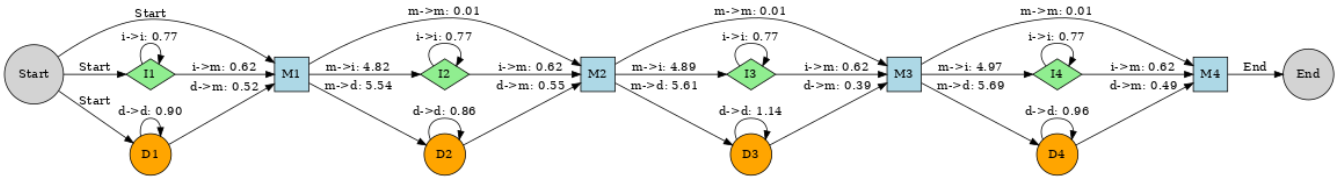

Fig.3. Graph representation of the Hidden Markov Model architecture for the first four states of the model trained in this paper. Transition and emission probabilities were taken from structural_model.hmm.

3. Results

3.1 Training sequences selection and HMM building

To build an efficient Hidden Markov Model it is important to have a very well-curated dataset, with high reliability on the sequence of the target domain. A total of 158 proteins were retrieved from Uniport and saved on a .csv file (*rcsb_pdb_custom_report_.csv*) supplement materials).
Structures sharing more than 95% of sequence identity were clustered together with CD-HIT.
After filtering procedures, from the 25 clusters (table 1) were chosen 24 representatives' proteins, the first cluster showed a single protein (2ODY_E), that was discarded ((*pdb_kunitz_customreportedfiltered. Fasta)*.
The multiple structure alignment resulted in 24 aligned residues, with an overall RMSD of 1.126 and an overall Q-score of 0.2318 (*pdb_kunitz_rp_formatted.ali)*.
Our Hidden Markov Model (*structural_model.hmm*) was constructed from 23 representatives' proteins. 5JBT, showing a RMSD higher than 2, was discarded.

Table 1. List of protein sequences from PBDefold MSA. 5JBT, showing a RMSD higher than 2, was discarded.

Number	PDB ID	Nres	Nsse	*RMSD*(Å)	*Q*-score
1	5nx1/C	54	4	0.4107	0.5090
2	6bx8/B	55	4	0.4000	0.5002
3	4bqd/A	55	4	0.3730	0.3535
4	4bx7/I	66	4	0.4330	0.4190
5	1yc0/B	66	4	0.4317	0.4355
6	5px5/A	66	4	0.4310	0.4356
7	1yc0/B	66	4	0.4357	0.4310
8	1f5r/A	66	4	0.4310	0.4455
9	1dt*R*/B	63	3	0.3114	0.4520
10	5jrol/A	57	4	0.4311	0.4455
11	3yyb/B	63	4	0.4310	0.4455
12	1knt/A	56	4	0.4517	0.4547
13	7byı/B	63	4	0.4510	0.4550
14	1zr0/X	56	4	0.4311	0.4530
15	6yyb/B	55	4	0.5027	0.4458
16	1knt/A	59	4	2.9166	0.4510
17	5ylv/A	56	3	2.1914	0.4455
18	1knt/A	55	4	0.4311	0.4560
19	5by7/A	59	4	0.5107	0.4455
20	5jb7/A	59	4	0.4311	0.4455
21	4uxb/A	56	4	0.4515	0.4132
22	6knt/A	55	4	0.4311	0.4550
23	3yyb/A	55	4	0.4527	0.4455
24	5jbt:/X	54	3	2.9166	0.4378

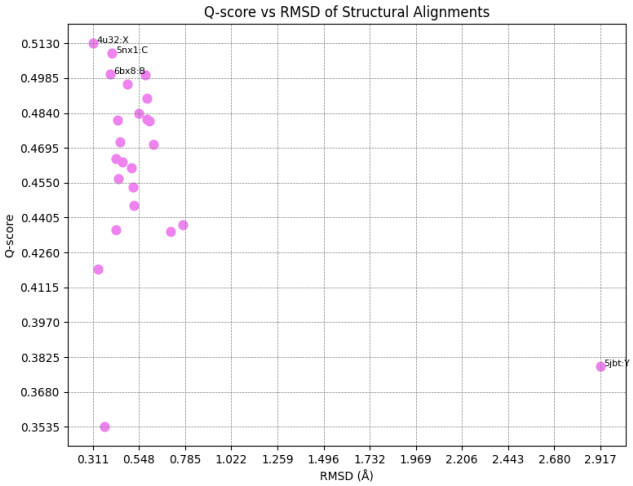

Fig. 4. Q-score and RMSD plot from MSA. 5JBT can be considered the outlier, with an RMSD of over 2.

Notably, our MSA shows that each one of our sequences had an initial unaligned part. This is given by the fact that, even if we are considering the same Kunitz domain, our total proteins are not identical, and our domain might be placed in different parts of our protein.

3.2 Test set generation and cleaning

To test our model, it was chosen a 2-k fold cross validation procedure, creating a positive and negative set, randomized and divided into two subsets(*neg_1.fasta*, *pos_1.fasta*, *neg_2.fasta*, *pos_2.fasta*). To avoid any biases, it was necessary to remove the ids of the proteins used for the HMM model creation. For this reason, a BLAST search against all the Kunitz proteins was performed. We filtered our output file using a specific python script (*get_seq.py*).

3.3 Model evaluation

Hmm search was used to find similar sequences using our negative and positives sets.
Results were saved and converted into a classification format. Positive and negative sets were merged for the first and the second set (*set_1.class*, *set_2.class*).

Table 2. Performance metrics with best e-value chosen

Set	E-val	MCC	TPR	PPV
Set 1	1e-06	0,9945	0,994	0,994
Set 2	1e-06	0,9863	0,983	0,980

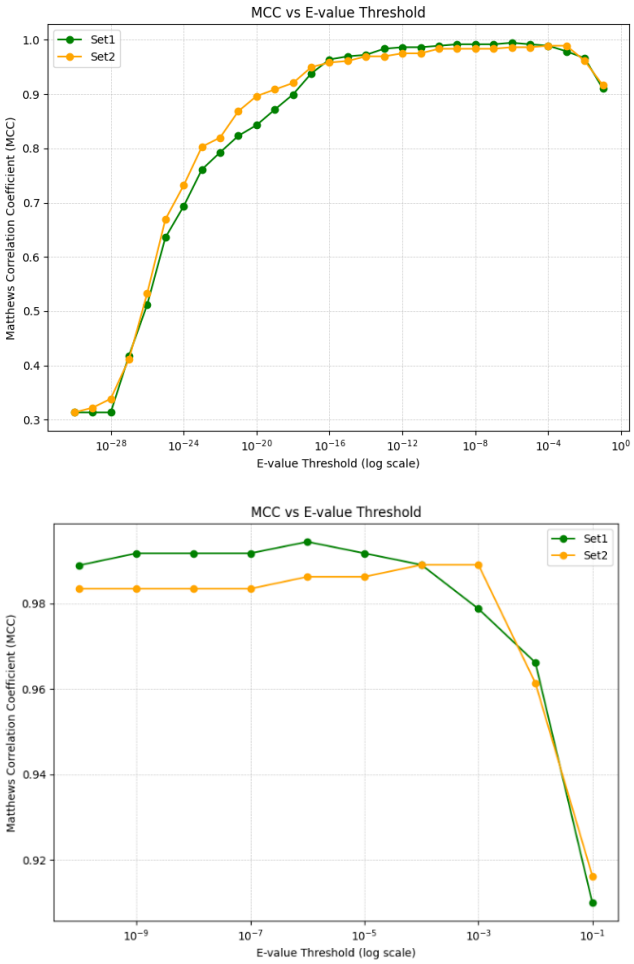

Fig. 5-6. Plot illustrating MCC values obtained using diverse E-value thresholds from 10-30 to 10-1, showing a clear relationship between MCC and E values.

A python script was created to evaluate the performance of our model (*performance.py*).
A range of e-values between 1e-30 and 1e-1 was tested on both subsets and for each threshold(figure 5-6), the accuracy, MCC, TPR, and FPR were computed(table 2). After averaging the thresholds that maximized the Matthews Correlation Coefficient, the selection of the optimal one was made. We identified optimal E-value thresholds for each iteration and we noticed a direct relationship between E-value thresholds and MCC: lower E-values corresponded to lower MCC values. Based on the MCC value, the best threshold chosen would be 1e-06(fig7-8). This important threshold maintains its statistical value, being small enough.
Overall, these results validate the reliability and generalization capability of the model in accurately identifying instances of interest within unseen data.

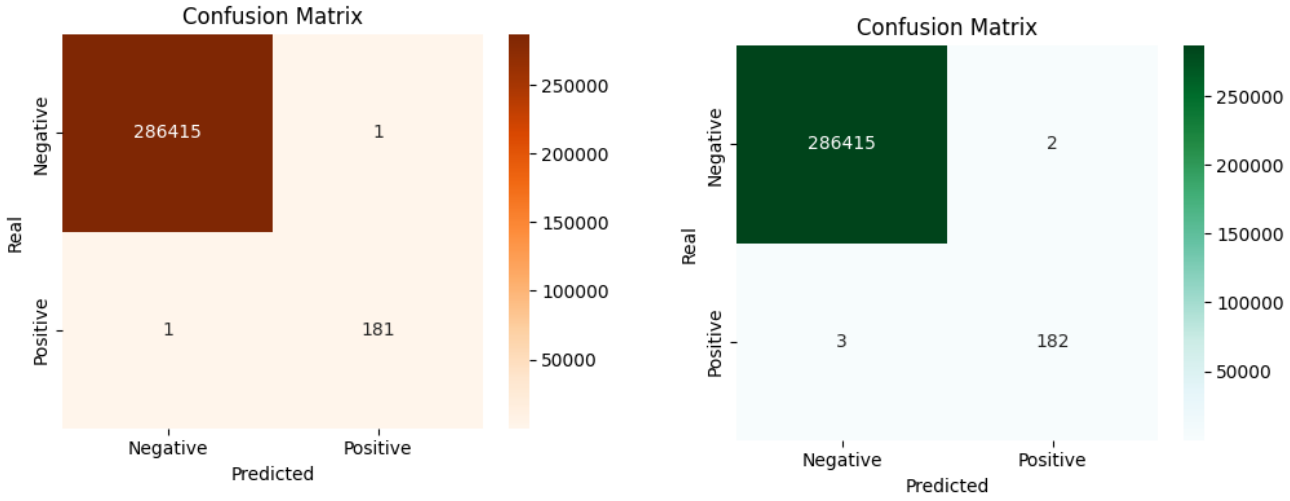

Fig.7-8. Confusion matrices for the final test on set_2 (orange) and set_1 (green), illustrating the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) at threshold 1e-06.

4. Discussion

In this project, we successfully designed and implemented a structure-informed bioinformatics pipeline for the identification of the Kunitz-type protease inhibitor domain using a Profile Hidden Markov Model (HMM). By integrating high-quality structural data, precise sequence filtering (via CD-HIT and BLAST), and alignment techniques (PDBeFold), we ensured the generation of a reliable and representative training set for model construction.

As we showed in the results, our HMM model (figure 3) accurately captured the conserved features of the Kunitz domain, including key disulfide bond-forming cysteines and functionally relevant motifs. The essential use of a 2-fold cross-validation approach, along with rigorous dataset partitioning and identity filtering, allowed for robust performance evaluation while minimizing the risk of overfitting.

Evaluation of the model on independent positive and negative datasets confirmed high classification performance across a range of E-value thresholds. Particularly, we chose the Matthews Correlation Coefficient (MCC), a balanced metric suitable for imbalanced data.

Minor misclassifications (e.g., sequences with high RMSD or potential misannotations) were handled carefully and highlighted the challenges of biological variation and the importance of accurate database curation. Overall, the model demonstrated strong reliability and can serve as an effective computational tool for detecting the Kunitz domain in large-scale sequence analysis pipelines.

Future improvements could include the integration of additional structural features, secondary structure predictions, or even machine learning-based ensemble methods to further refine classification, especially in borderline cases. Nonetheless, the current implementation lays a solid foundation for domain-specific HMM-based prediction workflows in functional annotation tasks.

References

1] E. Fratini et al. Molecular characterization of kunitz-type protease inhibitors from blister beetles
(coleoptera, meloidae). Biomolecules, 12:988, 2022.

[2] Kunitz domain - an overview. ScienceDirect Topics. https://www.sciencedirect.com/topics/
neuroscience/kunitz-domain.

[3] H. Zhang et al. Bioinformatic comparison of kunitz protease inhibitors in echinococcus granulosus sensu
stricto and e. multilocularis and the genes expressed in different developmental stages of e. granulosus
s.s. BMC Genomics, 22:907, 2021.

[4] S Ranasinghe and DP McManus. Structure and function of invertebrate kunitz serine protease inhibitors.
Developmental & Comparative Immunology, 39(3):219–227, Mar 2013. Epub 2012 Nov 24.

[5] K. Jablonowski. Hidden markov models for protein domain homology identification and analysis. In
K. Machida and B. A. Liu, editors, SH2 Domains: Methods and Protocols, pages 47–58. Springer, New
York, NY, 2017.

[6] B.J. Yoon. Hidden markov models and their applications in biological sequence analysis. Current
Genomics, 10(6):402–415, Sep 2009.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E.
Bourne. The protein data bank. Nucleic Acids Research, 28(1):235–242, 2000.

[8] Eugene Krissinel and Kim Henrick. Secondary-structure matching (ssm), a new tool for fast protein
structure alignment in three dimensions. Acta Cryst. D, 60:2256–2268, 2004.

[9] Martin Larralde et al. Pyhmmer: a python library binding to hmmer. 2022.
[10] Sean R Eddy. Hmmer3: a new generation of sequence homology search software. Bioinformatics,
27(17):2957–2958, 2011.
[11] Travis J Wheeler, Jody Clements, and Robert D Finn. Skylign: a tool for creating informative, interactive
logos representing sequence alignments and profile hidden markov models. BMC Bioinformatics, 15(1):7,
2014.

[12] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res.,
51:D523–D531, 2023.
[13] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local
alignment search tool. Journal of molecular biology, 215:403–410, 1990.