RBP-ID – Development of a Web-Based Bioinformatics Tool for the Identification of Phage Receptor-Binding Proteins

Cátia Rosário[1]

[1] Minho University, R. da Universidade 4710-057, PORTUGAL
PG57791@alunos.uminho.pt[1]

Abstract. Bacteriophages are viruses that infect bacteria, playing a crucial role in numerous areas. Their host specificity is determined by receptor-binding proteins (RBPs), which intervene in bacterial detection and attachment. Identifying RBPs is challenging due to their sequence diversity and low conservation among phages. To address this, we intend to create RBP-ID, a web-based bioinformatics tool for systematic RBP prediction in phage genomes. By leveraging validated and predicted RBP datasets, RBP-ID will integrate multiple computational approaches in a sequential workflow. The methodology will first involve sequence similarity assessment using CD-HIT, followed by the application of unsupervised machine learning like K-means for RBP clustering. The classification will then be refined using supervised machine learning like KNN to measure similarity. This process aims to enhance annotation accuracy and aid in novel RBP discovery. The tool aspires to provide ranked lists of predicted RBPs with detailed annotations and similarity scores in a user-friendly interface. This tool represents a significant step toward improved phage genome annotation and the biotechnological potential of RBPs in targeted phage therapy, biosensors, and microbiome engineering.

Keywords: RBP, Receptor-Binding Protein, Bacteriophages, Phages, Fibers, Tailspikes, Baseplates, Database, Machine-Learning.

1 Introduction

Bacteriophages are viruses also known as phages that specifically infect bacteria. They play a critical part regulating the populations of bacteria and by shaping community dynamics of microbes in environments such as marine ecosystems, soil, and the human microbiome [1-3,9,11,60]. Phages are recognized for their biotechnological and therapeutic potential, particularly in addressing antibiotic-resistant infections and microbiome engineering [4,18-22]. Their ability to precisely target pathogenic bacteria without harming beneficial microbes makes them an attractive alternative to broad-spectrum antibiotics, especially in the context of antimicrobial resistance (AMR). Moreover, through co-evolutionary processes, phages can adapt to bacterial mechanisms, including receptor mutations and CRISPR-Cas defenses, sustaining their therapeutic efficacy [8-10,20].

Advances in systems and synthetic biology have significantly broadened the scope of phage applications beyond conventional therapy. Engineered phages and phage-derived enzymes, such as endolysins and depolymerases, are being tailored for diverse clinical, agricultural, and industrial contexts. Furthermore, phage display technologies and phage-based delivery systems have emerged as sophisticated platforms for targeted gene editing and the delivery of therapeutic biomolecules [15–19,23].

A critical determinant of these expanded applications lies in the phage's capacity to specifically recognize and infect bacterial hosts, a process that's mediated by receptor-binding proteins (RBPs). RBPs facilitate the initial interaction with bacterial surface structures - including lipopolysaccharides, teichoic acids, and capsular poly-saccharides - thereby dictating host specificity and infection efficiency [2,19]. These proteins are typically localized on the tail fibers, tail spikes, or baseplate of phages, and their structural heterogeneity reflects the co-evolutionary arms-race among bacteriophages and their respective hosts [11,18,55]. RBPs may be presented as elongated tail fibers or compact tail spike proteins (TSPs), which are multidomain proteins involved in host receptor recognition [34,40,49].

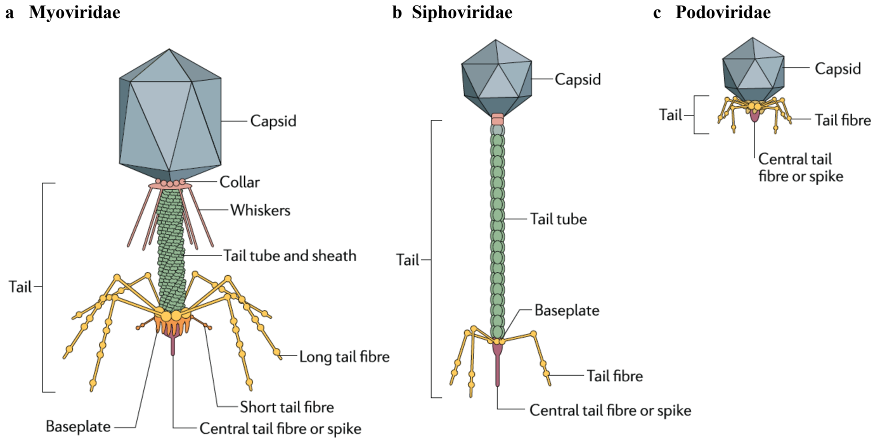


Fig. 1. Structures of the three tailed morphotypes of the order *Caudovirales*. (a) Myoviridae is the only family with a contractile tail. (b) Siphoviridae possess at the end of the tail a baseplate, like Myoviridae, yet has a non-contractile tail. (c) Podoviridae is the only family without a baseplate and with a short non-contractile tail (Nobrega *et al.*, 2018) [51,60].

Structural variation in RBPs contributes to the specificity of different phage morphotypes, such as Myoviridae**,** Siphoviridae, and Podoviridae, all members of the class of tailed viruses *Caudoviricetes*. Myoviridae is characterized by contractile tail fibers and a complex baseplate which employs a multi-component system for bacterial adhesion and penetration. Siphoviridae have long, non-contractile tails and rely on tail fibers to recognize host receptors, whereas Podoviridae, with their short, non-contractile tails, use adhesion proteins at the tail base to initiate infection [11,13,24,51] (Fig.1).

Following adsorption, phages deploy structural adaptations to breach bacterial defenses, with contractile-tailed phages using sheath contraction and non-contractile-tailed phages relying on enzymatic degradation of bacterial cell walls [15-19]. Once inside the host, phage genomes hijack the bacterial cellular machinery to either initiate a lytic cycle, culminating in bacterial lysis and phage progeny release, or a lysogenic cycle, resulting in prophage integration, which influences their potential applications in various domains, including biocontrol and microbiome modulation [2, 8-12,20].

In addition to mediating host specificity, RBPs are key targets for phage engineering with altered ranges, enabling synthetic phages creation for personalized therapy. RBPs also have emerging applications in biosensing and microbiome engineering, where they facilitate the detection or selective modulation of bacterial populations in clinical, environmental, and agricultural settings [7,9–11,17–20,56].

Despite their importance, identifying RBPs remains challenging due to their significant sequence diversity, even among closely related phages. This variability limits the effectiveness of conventional annotation pipelines and makes functional characterization difficult [13,20].

Traditional methods like BLAST are effective for finding closely related sequences but struggle to identify RBPs with divergent sequences. Hidden Markov Models (HMMs), employed by tools such as HMMER, improve sensitivity, but they require well-curated training dataset [5-7,25-27].

PhANNs offer a high-throughput, deep learning-based alternative to traditional sequence-based methods, designed to classify structural phage proteins rapidly, reducing computation time while maintaining high exactness. Although methods like these are becoming increasingly relevant for genome-wide RBP annotation, offering efficient and reliable identification of RBP sequences, they are still in development and face limitations due to the lack of comprehensive datasets for training models [5,9,18,48].

Boeckaerts et al. combined domain-based searches using HMMs with machine learning classifiers such as Extreme Gradient Boosting (XGBoost). HMMs were used to detect conserved domains in RBP sequences, while XGBoost, trained on known RBP sequences, was able to distinguish RBPs from other phage proteins. This hybrid approach improved the precision of RBP functional annotation [5].

The employment of Random Forest and Support Vector Machines (SVMs) - Machine learning (ML) classifiers -, to predict bacterial hosts based on the RBP composition. By analyzing annotated RBP sequences from databases like PhagesDB and NCBI RefSeq, these classifiers can predict the bacterial host of a given phage, representing a promising direction for targeted phage therapy and biocontrol applications [6,44].

PHYPred is an example of a functional annotation tool that focuses on predicting phage-encoded enzymes involved in bacterial cell wall degradation. While initially designed for hydrolases and endolysins, the methodology used can be adapted for RBP identification. The integration of domain-specific databases like MEROPS and CAZy into predictive models enhances the accuracy of these annotations [8].

Dunne et al. introduced a structure-guided methodology for modifying RBPs to reprogram phage host specificity. This approach leverages structural insights from crystallography to identify key residues responsible for receptor binding. Tools like AlphaFold for prediction protein structure and PyMOL for molecular visualization have been instrumental in understanding RBP function, enabling the engineering of chimeric RBPs for new bacterial host recognition [7].

While computational tools provide valuable insights, experimental validation remains crucial. Simpson et al. proposed an experimental essay for isolating and characterizing RBPs, which can be used to validate computational predictions. Tools like Clustal Omega for sequence alignment and TMHMM for transmembrane domain prediction are frequently employed to confirm predictions and further refine the functional annotations of RBPs [3].

To address the challenges described, an alternative approach involves focusing on conserved regions within RBPs that are critical for stability and host interaction. Studies suggest that the N-terminal region of RBPs, particularly in tail fibers, tail spikes, and baseplate proteins, exhibits a higher degree of conservation than other regions. This conserved architecture is essential for protein folding and proper assembly within the phage structure, making it a promising target for bioinformatics methods aimed at identifying RBPs [17-19]. Exploiting these conserved domains, along with an integrative approach combining comparative genomics such as Diamond, HMM-based domain searches, unsupervised and supervised machine learning algorithm for classification, like K-means and K-Nearest Neighbors (KNN), respectively, can improve the accuracy of RBP identification predictions [5,28, 41,42,54].

A specialized computational tool dedicated to RBP identification would provide a significant advancement in the annotation of phage genomes. By integrating multiple analytical approaches, such a tool could facilitate the rapid detection of RBPs, thereby accelerating the functional characterization of phage genes and proteins [5,18,29]. This capability would support the various applications mentioned before, including the development of phage-based biocontrol strategies, and the refinement of microbiome modulation techniques. Additionally, a robust bioinformatics pipeline for RBP identification and annotation would be a valuable resource for researchers and clinicians, once it would enhance our understanding of phage-host interactions [10,18,30].

2 Methodology

2.1 Dataset Collection and Filtering

To ensure the creation of a comprehensive database for receptor-binding proteins identification, three distinct datasets were compiled separately: one consisting of experimentally validated phage RBPs, other containing computationally predicted phage RBPs and another that combined the previous two.

2.1.1 Validated and Computed RBP Dataset

For experimentally validated RBPs, data was gathered from both existing literature and publicly available sources, including UniProtKB, NCBI, and PDB. Specific filtering methods were applied to each source to ensure only relevant entries were collected:

- **UniProtKB**: Search using "phage tail fiber" and restricted to Caudoviricetes (taxonomy_id: 2731619). Additional filters to refine results to reviewed entries were applied [32,51].

QUERY: (reviewed:true) **AND** (keyword:KW-1230) **AND** (taxonomy_id:2731619)

- **NCBI**: A broad search for RBPs was conducted, restricted to Caudoviricetes, and filtered to retain only one representative protein per RBP, avoiding redundancy [33,51].

QUERY: (tail fiber) **AND** "Caudoviricetes sp."[porgn:__txid2832643]

- **PDB**: Keywords such as "tail fiber" and "baseplate" were used to find annotated RBP proteins. Filters ensured relevance to Caudoviricetes and only experimentally validated structures were included [2,10,50,51].

QUERY: Source Organism Taxonomy Name (Full Lineage) = "Caudoviricetes" **AND** (Structure Title **HAS EXACT PHRASE** "tail fiber" **OR** Structure Title **HAS EXACT PHRASE** "baseplate") **AND** Experimental Method **EXISTS**

The structure of the first dataset included 13 variables defined a priori namely: Data Origen, Phage Accession nr, Bacteriophage, Taxonomic Classification, Morphotype, Host, Complete Phage Name, Phage Deliniation, RBP Name, RBP Accession, Protein Type, Annotation, RBP Sequence (aa), Delineation (aa), Expression System.

Alongside the experimentally validated RBPs, a second dataset was created, containing computationally predicted RBPs, obtained from literature sources [6]. To ensure consistency, this dataset was structured to match most of the variables of the validated RBP dataset.

After compilation, to facilitate computational analysis, the datasets were processed with Biopython. To remove duplicates and perform K-means, the CleanSeq column with sequences without FASTA header (maintaining the duplicated IDs in a new column to not lose information) was created. Another column called FASTA was created with the sequences with FASTA header (if missing, headers were created) and used to create the three FASTA files to further along be used in CD-HIT analysis [42,54,55].

2.1.2 Homology-Based Identification and Clustering

This phase of the project employed a structured, multi-stage methodology for the analysis of Receptor-Binding Proteins (RBPs).

Firstly, CD-HIT was used for sequence similarity and redundancy reduction at a 0.4 identity threshold, yielding representative sequences [42]. Both original and representative RBP sequences were then vectorized using CountVectorizer (3-4 mers). K-means clustering, an unsupervised method, was applied to these vectorized sequences, with optimal K determined by the Silhouette Score (k=2-10) [54, 55]. Cluster assignments were mapped to morphotypes via majority vote. Principal Component Analysis (PCA) was used for visual exploration of clusters in reduced dimensions. Hierarchical Dendrograms, employing cosine distance and Ward's method, were constructed to visualize clustering for all RBPs, validated RBPs, and non-validated RBPs (truncated at level 6, p=6, leaf_rotation=90). The preliminary K-means classification performance was evaluated with a Confusion Matrix and Classification Report values (Recall, F1-score, Precision) [58]. Future work will involve refining classification with k-Nearest Neighbors (KNN) and validating results through statistical evaluations, ultimately contributing to a systematic RBP identification and functional annotation tool [38, 41].

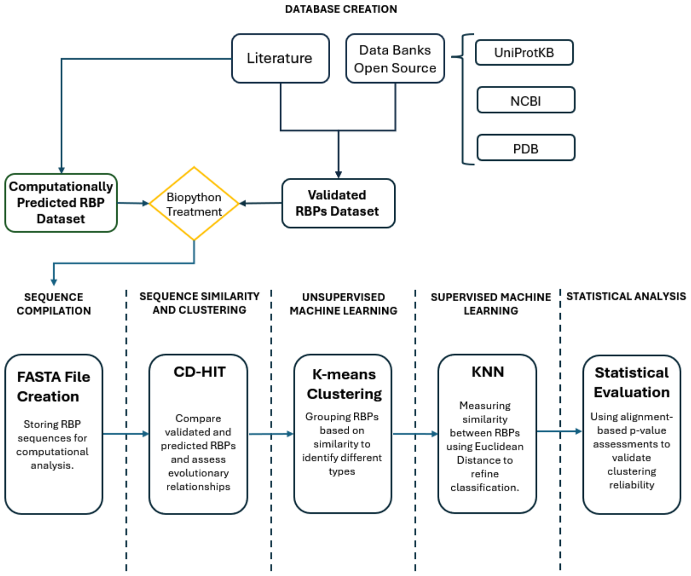

DATABASE CREATION
Literature
Data Banks Open Source
UniProtKB
NCBI
PDB
Computationally Predicted RBP Dataset
Biopython Treatment
Validated RBPs Dataset
SEQUENCE COMPILATION
SEQUENCE SIMILARITY AND CLUSTERING
UNSUPERVISED MACHINE LEARNING
SUPERVISED MACHINE LEARNING
STATISTICAL ANALYSIS
FASTA File Creation
Storing RBP sequences for computational analysis.
CD-HIT
Compare validated and predicted RBPs and assess evolutionary relationships
K-means Clustering
Grouping RBPs based on similarity to identify different types
KNN
Measuring similarity between RBPs using Euclidean Distance to refine classification.
Statistical Evaluation
Using alignment-based p-value assessments to validate clustering reliability

Fig. 2. Workflow for the Creation and Analysis of a Phage RBP Dataset. Project began with a compilation of two RBP datasets: a validated set derived from literature and open-source data banks (UniProtKB, NCBI, PDB), and a computationally predicted set. The main analytical pipeline starts with formatting all sequences into a FASTA file. Sequence similarity is then assessed using CD-HIT, followed by unsupervised clustering with K-means to identify RBP types. The classification is refined using the supervised KNN algorithm. Finally, the reliability of the clustering is confirmed through statistical evaluation using alignment-based p-value assessments [38-42,53-59,63].

3 RESULTS AND DISCUSSION

Building upon the outlined methodology, this section presents the results obtained, which are discussed in the context of RBP characterization and prediction.

The confusion matrix and classification report (Annexes Fig.1) offer insights into the performance of the model classifying "Myoviridae," "Podoviridae," and "Siphoviridae" samples [62]. The matrix reveals a significant challenge in distinguishing between "Myoviridae" and "Siphoviridae," as 14 actual "Myoviridae" samples were incorrectly predicted as "Siphoviridae," and 16 actual "Podoviridae" samples were also misclassified as "Siphoviridae." This is further supported by the classification report, where "Myoviridae" shows very low recall (0.07), indicating that the model struggles to identify most of the actual "Myoviridae" instances. Similarly, "Podoviridae" has a moderate recall of 0.36, suggesting a substantial number of false negatives for this class. In contrast, "Siphoviridae" demonstrates excellent recall (1.00), meaning all actual "Siphoviridae" samples were correctly identified. However, its precision (0.38) is low, implying that a large proportion of instances predicted as "Siphoviridae" were from other classes, specifically "Myoviridae" and "Podoviridae," as seen in the confusion matrix. The F1-scores reflect these trends, with "Myoviridae" having a very low F1-score (0.12), "Podoviridae" a moderate F1-score (0.53), and "Siphoviridae" a slightly higher F1-score (0.55) despite its low precision, primarily due to its perfect recall. Overall, the model appears to be biased towards predicting "Siphoviridae," leading to good recall for that class but poor precision and significantly hindering the accurate classification of "Myoviridae" and "Podoviridae."

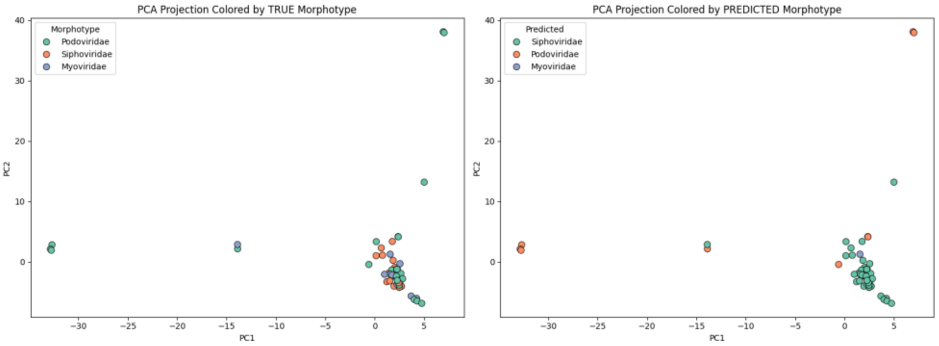


Fig. 3. PCA Projections of Data. The left plot displays the data projected onto PC1 and PC2, and points are colored according to their TRUE morphotype (Podoviridae, Siphoviridae, Myoviridae). The right plot shows the same PCA projection, but with points colored according to their Predicted morphotype by the classification model [63].

In the Figure 3, PCA plots provide a visual comparison of the true morphotype distribution and the model's predicted morphotype distribution in a lower-dimensional space. In the graph on the left, it's observed that "Podoviridae" samples are largely clustered in a distinct group around PC1 = -30, PC2 = 2. The "Siphoviridae" (orange) and "Myoviridae" (blue) morphotypes, however, show significant overlap and intermixing within the region of PC1 values between -5 and 5, with some outliers for both "Siphoviridae" and "Myoviridae" appearing at higher PC2 values (e.g., PC2 around 35-40). This inherent overlap between "Siphoviridae" and "Myoviridae" in the true data suggests that these classes are not linearly separable based on these principal components, which could pose a challenge for classification [63].

Turning to the graph on the right, the model appears to correctly identify the distinct cluster of "Podoviridae" (teal) around PC1 = -30, PC2 = 2, as this cluster remains largely teal in the predicted plot. However, for the more intermixed "Siphoviridae" and "Myoviridae" regions, the predictions show considerable misclassification, particularly for "Myoviridae." While some "Siphoviridae" (orange) predictions align with the true Siphoviridae cluster, there's a strong tendency for the model to predict "Siphoviridae" in areas where true "Myoviridae" are located. This aligns with findings from a confusion matrix (Fig.3), confirming the high recall for "Siphoviridae" but low precision, and conversely, low recall for "Myoviridae" due to them being misclassified as "Siphoviridae." The outlier "Myoviridae" point at PC1 ~5, PC2 ~38 in the true plot is also predicted as "Siphoviridae," further highlighting the model's difficulty in distinguishing these two morphotypes when they are truly "Myoviridae" [63].

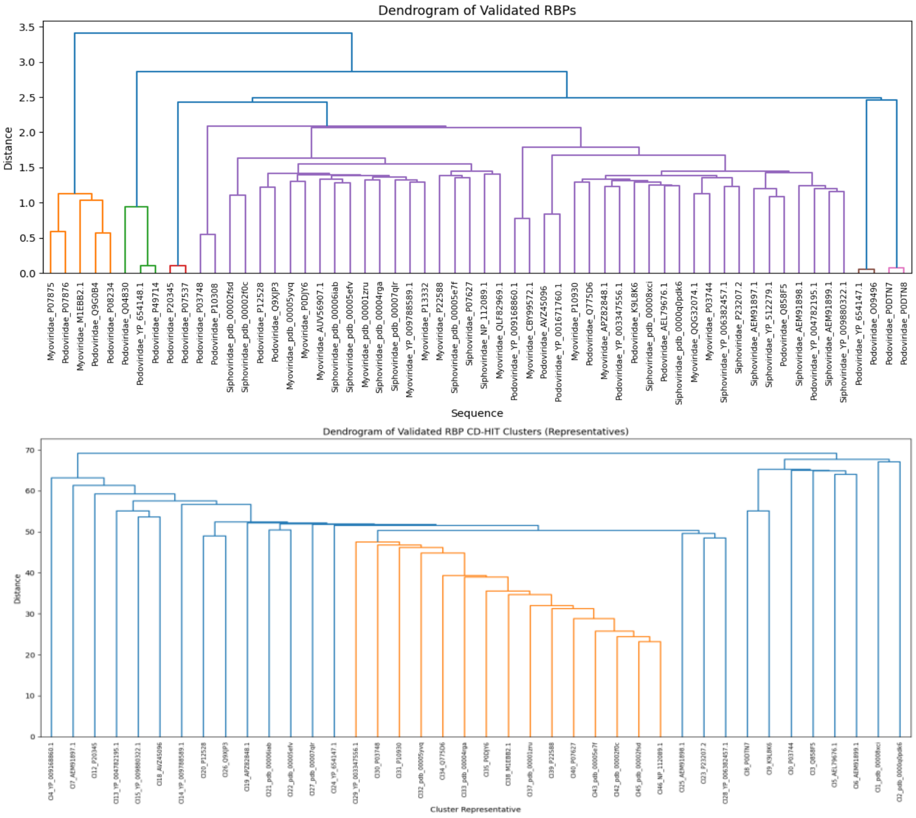


Fig. 4. Dendrograms of Validated RBPs. The top dendrogram illustrates hierarchical clustering of individual RBP sequences, vectorized using CountVectorizer (3-4 mers) and clustered with Ward's method based on cosine distance. The bottom dendrogram shows hierarchical clustering of representative sequences obtained from CD-HIT clusters (with a sequence identity threshold of 0.3), using the same vectorization and clustering parameters [42,53,54,64].

Both dendrograms in Figure 4 depict hierarchical clustering of RBP sequences, offering insights into their similarity and grouping patterns, albeit derived from different starting points. The top dendrogram, representing individual validated RBPs, reveals distinct clusters, particularly for "Myoviridae" and "Podoviridae" sequences, which tend to form compact groups at lower distances. However, "Siphoviridae" sequences appear more spread out and intermingled with others, suggesting higher diversity within this morphotype or less distinct sequence features. This dendrogram is based on direct sequence vectorization and cosine distance, providing a granular view of relatedness.

In contrast, the bottom dendrogram, derived from CD-HIT representative sequences (clustered at a 0.3 sequence identity threshold), presents a more abstract view of clustering by first reducing redundancy. The clusters formed in this dendrogram likely represent broader groupings of highly similar sequences. While the specific labels differ, the overall structure of the bottom dendrogram, with its larger, more broadly defined branches, implies that CD-HIT effectively grouped highly similar sequences, and the subsequent clustering of these representatives highlights the higher-level relationships between these pre-clustered groups [42,53,54,64].

Comparing the two, the top dendrogram offers fine-grained details of individual sequence relationships, while the bottom dendrogram provides a more generalized understanding of the structural relationships between the major RBP sequence families after redundancy reduction. The difference in vertical scale between the two dendrograms also suggests varying levels of dissimilarity being represented, with the CD-HIT representatives showing larger overall distances, likely due to the inherent differences between the chosen cluster representatives [42,53,54,64].

3.1 RBP NON-VALIDATED AND BOTH VALIDATED AND NON-VALIDATED DATASETS

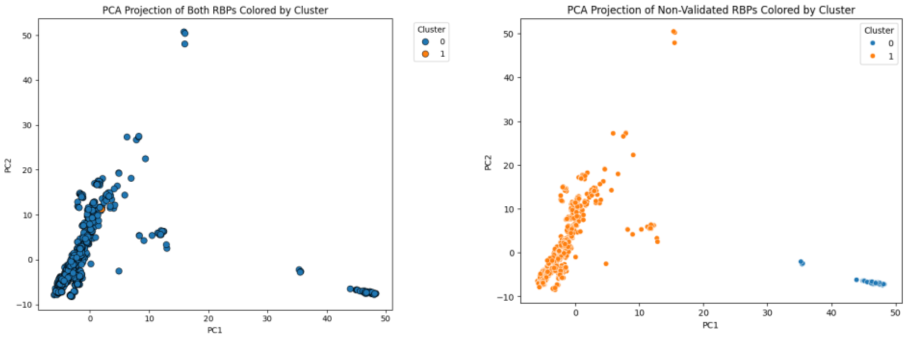


Fig. 5. PCA Projections of RBP Data Colored by Cluster. The left plot shows the PCA projection of both validated and non-validated RBPs, with points colored according to their assigned cluster (0 or 1). The right plot displays the PCA projection exclusively for non-validated RBPs, also colored by cluster [63].

These PCA plots illustrated in Figure 5, the clustering results within a two-dimensional principal component space, allowing for a visual assessment of how different RBP groups are separated. The left plot reveals a predominant cluster (Cluster 0, blue) that forms a dense grouping from approximately PC1 = -10 to 10 and PC2 = -10 to 30. This large cluster suggests a significant proportion of the RBPs share similar characteristics in this transformed space. Within this dominant blue cluster, a smaller number of points belonging to Cluster 1 (orange) are interspersed, indicating some overlap in features within this central region. Notably, there's also a distinct sub-cluster of Cluster 0 (blue) points observed further along the PC1 axis, specifically around PC1 values of

40-50 and PC2 values of -5, indicating a separate group of RBPs with very different characteristics [63].

The right plot provides further insight into the composition of these clusters. In this plot, Cluster 1 (orange) is overwhelmingly dominant, forming a large and somewhat diffuse cluster that largely overlaps with the central dense region of Cluster 0 from the "Both RBPs" plot (PC1 = -10 to 10, PC2 = -10 to 30). This suggests that the majority of non-validated RBPs fall into Cluster 1. A few points belonging to Cluster 0 (blue) are also present in the non-validated set, located in the distinct region around PC1 = 40-50, PC2 = -5. This distribution strongly implies that the non-validated RBPs largely constitute the Cluster 1 group, which in the combined dataset is somewhat intermingled with the main Cluster 0, while the specific sub-cluster at high PC1 values remains predominantly Cluster 0, irrespective of validation status for those few non-validated points present there [63].

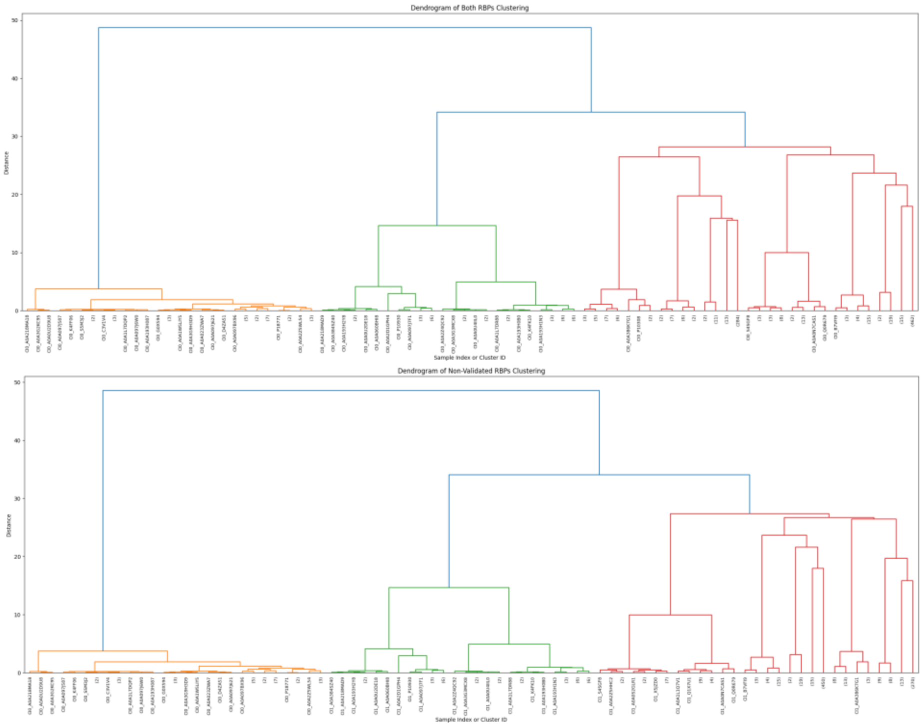


Fig. 6. Dendrograms of K-means Clustering on RBP Sequences. The top dendrogram illustrates the hierarchical clustering of both validated and non-validated RBP sequences after initial K-means clustering (optimal K determined by silhouette score for 3-4 mer CountVectorizer features), using cosine distance and Ward's method. The bottom dendrogram displays the same hierarchical clustering process applied exclusively to non-validated RBP sequences. Both dendrograms are truncated at level 6 with leaf labels rotated by 90 degrees [42,53,54,64].

The dendrograms in Fig. 7 depict the hierarchical relationships of RBP sequences after an initial K-means clustering step based on k-mer CountVectorizer features. The top dendrogram shows three primary clusters (distinguished by orange, green, and red branches) at higher distances. The large blue branch at the top suggests a broad overall grouping. Interestingly, the K-means approach appears to have segregated sequences into distinct initial groups that are then further hierarchically clustered. The left-most orange cluster contains a few tightly grouped sequences, while the green and red clusters are more diverse in their internal structure, showing sub-clusters at varying distances. The bottom dendrogram, focused solely on "Non-Validated RBPs" mirrors the general structure observed in the "Both RBPs" dendrogram. This similarity suggests that the non-validated sequences distribute themselves across these established clusters in a manner consistent with their contribution to the overall RBP population. Both dendrograms indicate a hierarchical structure where sequences form tight sub-clusters at lower distances before merging into broader groups, reflecting varying degrees of sequence similarity or shared functional motifs captured by vectorization. The truncation at p=6 allows for a view of the major branching patterns without overwhelming detail [42,53,54,64].

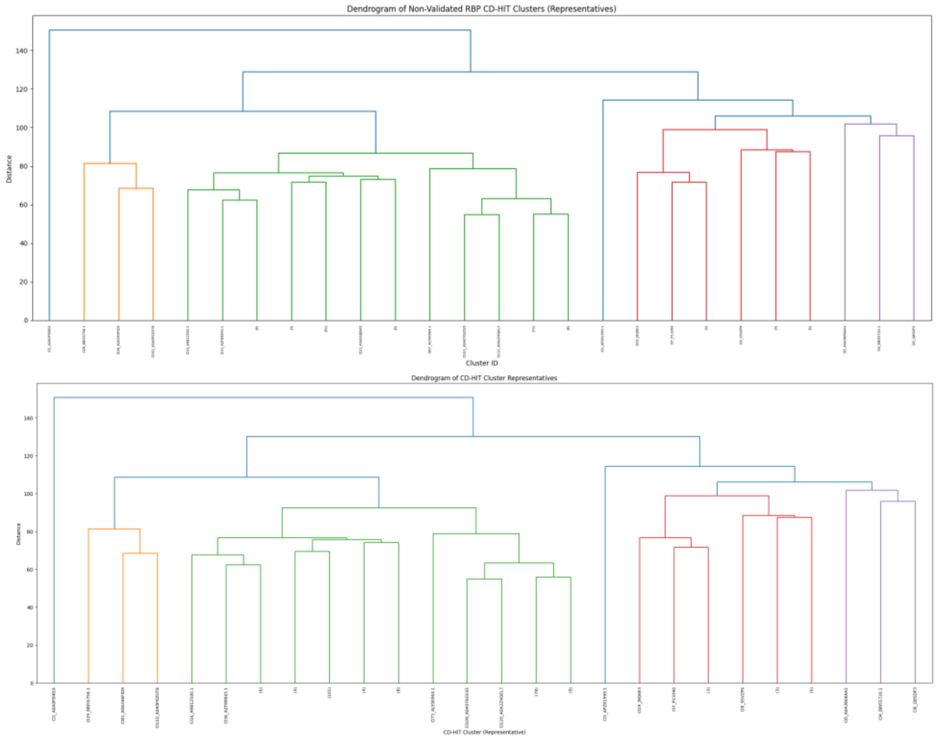


Fig. 7. Dendrograms of Hierarchical Clustering on CD-HIT Cluster Representatives. The top dendrogram shows the hierarchical clustering of representative sequences from CD-HIT clusters (at a 0.4 sequence identity threshold) derived from all RBPs, using 3-4 mer CountVectorizer features, cosine distance, and Ward's method. The bottom dendrogram applies the same

clustering methodology to CD-HIT cluster representatives specifically from non-validated RBPs. Both dendrograms are truncated at level 6 with leaf labels rotated by 90 degrees [42,53,54,64].

The dendrograms in Figure 7 illustrate the hierarchical relationships among representative sequences generated by CD-HIT (with a 0.4 identity threshold), offering a high-level view of RBP family structures. The top dendrogram, representing from all RBPs, clearly delineates distinct clusters (orange, green, red, and purple) at significant distances, implying substantial divergence between these major RBP families. The long vertical lines before merging indicate significant dissimilarity between these broad groups. The green cluster appears to be the most diverse internally, with several subbranches. The bottom dendrogram exhibits a striking resemblance in its overall structure and branching patterns to the top dendrogram. The presence of similar large clusters at comparable distances suggests that the non-validated RBPs largely contribute to and follow the same fundamental structural and sequence-based family distinctions as the full RBP dataset. The consistent large distances between the main branches in both dendrograms underscore that even after reducing redundancy with CD-HIT, the identified RBP families remain quite distinct based on their 3-4 mer compositions. The truncation helps to highlight these major cluster relationships [42,53,54,64].

4 Conclusions

The work presented marks a foundational step towards the development of RBP-ID, a comprehensive web-based tool for systematic RBP prediction. Through the application of unsupervised ML techniques, particularly K-means and hierarchical clustering on both raw and CD-HIT-reduced sequence data, we gained critical insights into the inherent structural and compositional relationships within RBP datasets, encompassing both validated and non-validated sequences. The clustering analyses effectively elucidated distinct RBP families and demonstrated that non-validated sequences largely conform to these established groupings, reinforcing their potential relevance [42,63,64].

While these initial explorations provided valuable insights into sequence relationships and the challenges in morphotype classification, they simultaneously highlighted the necessity for further refinement. The current unsupervised approaches, while powerful for discerning underlying patterns, require validation with robust classification.

Future work will critically involve the integration of supervised machine learning, specifically K-Nearest Neighbors (KNN), to enhance the predictive accuracy of RBP identification and to measure sequence similarity with greater precision. Crucially, a complete statistical assessment of the performance of the model will be undertaken, rigorously validating its efficacy. The culmination of these efforts will be the construction of the user-friendly web interface, RBP-ID, which will transform these analytical insights into a practical bioinformatics resource for phage genome annotation. This progression is essential for advancing the discovery of novel RBPs and unlocking their significant biotechnological potential in areas such as targeted phage therapy, biosensors, and microbiome engineering [41,57].

References

1. Zhang, Z., Yu, F., Zou, Y., Qiu, Y., Wu, A., Jiang, T., Peng, Y. Phage protein receptors have multiple interaction partners and high expressions, *Bioinformatics*, 36(10), 2975–2979, (2020). https://doi.org/10.1093/bioinformatics/btaa123
2. Pas, C., Latka, A., Fieseler, L. Briers, Y. Phage tailspike modularity and horizontal gene transfer reveals specificity towards *E. coli* O-antigen serogroups. *Virol J*, 20, 174, (2023). https://doi.org/10.1186/s12985-023-02138-4
3. Simpson, D.J., Sacher, J.C., Szymanski, C.M., Development of an Assay for the Identification of Receptor Binding Proteins from Bacteriophages. *Viruses*, 8(1):17, (2016) https://doi.org/10.3390/v8010017
4. Dufloo, J., Andreu-Moreno, I., Moreno-García, J. Valero-Rello, A., Sanjuán, R.. Receptor-binding proteins from animal viruses are broadly compatible with human cell entry factors. *Nat Microbiol*, 10, 405–419, (2025). https://doi.org/10.1038/s41564-024-01879-4
5. Boeckaerts, D., Stock, M., De Baets, B., Briers, Y. Identification of Phage Receptor-Binding Protein Sequences with Hidden Markov Models and an Extreme Gradient Boosting Classifier, *Viruses*, 14(6):1329, (2022). https://doi:10.3390/v14061329
6. Boeckaerts, D., Stock, M., Criel, B., Gerstmans, H., De Baets, B., Briers, Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep. 2021;11(1):1467, (2021). https://doi:10.1038/s41598-021-81063-4
7. Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M. J., Kilcher, S. Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Rep. 29(5):1336-1350.e4. (2019). https://doi:10.1016/j.celrep.2019.09.062.
8. Ding, H., Yang, W., Tang, H. Feng, P.M, Huang, J., Chen, W., Lin, H. *PHYPred*: a tool for identifying bacteriophage enzymes and hydrolases. *Virol. Sin.* 31, 350–352 (2016). https://doi.org/10.1007/s12250-016-3740-6
9. Cantu, V.A., Salamon, P., Seguritan, V., Redfield, J., Salamon, D., Edwards, R.A., Segall, A.M. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLoS Comput Biol*, 16(11): e1007845. (2020). https://doi:10.1371/journal.pcbi.1007845
10. Bull, J.J.,Vimr, E.R., Molineux, I.J. A tale of tails: Sialidase is key to success in a model of phage therapy against K1-capsulated Escherichia coli. *Virology*. 398. 1. 79-86. (2010). https://doi.org/10.1016/j.virol.2009.11.040.
11. d'Acapito, A., Roret, T., Zarkadas, E., Mocaër, P., Lelchat, F., Baudoux, A., Schoehn, G., Neumann, E.Structural Study of the Cobetia marina Bacteriophage 1 (Carin-1) by Cryo-EM. J Virol97:e00248-23. (2023). https://doi.org/10.1128/jvi.00248-23
12. Van den Berg, B., Silale, A., Baslé, A., Brandner, A.F., Mader, S.L., Khalid, S. Structural basis for host recognition and superinfection exclusion by bacteriophage T5, *Proc. Natl. Acad. Sci. U.S.A.* 119 (42) e2211672119. (2022). https://doi.org/10.1073/pnas.2211672119
13. Siponen, M., Spinelli, S., Blangy, S., Moineau, S., Cambillau, C., Campanacci, V. Crystal Structure of a Chimeric Receptor Binding Protein Constructed from Two Lactococcal Phages. J Bacteriol191. (2009). https://doi.org/10.1128/jb.01637-08
14. Bebeacua, C., Bron, P., Lai, L., Vegge, C.S., Brøndsted, L., Spinelli, S., Campanacci, V., Veesler, D., van Heel, M., Cambillau, C.: Structure and Molecular Assignment of Lactococcal Phage TP901-1 Baseplate. *J. Biol. Chem.* 285, 39079–39086 (2010). https://doi.org/10.1074/jbc.M110.175646
15. Legrand, P., Collins, B., Blangy, S., Murphy, J., Spinelli, S., Gutierrez, C., Richet, N., Kellenberger, C., Desmyter, A., Mahony, J., van Sinderen, D., Cambillau, C.The Atomic Structure of the Phage Tuc2009 Baseplate Tripod Suggests that Host Recognition Involves Two Different Carbohydrate Binding Modules. *mBio* 7:10.1128/mbio.01781-15. (2016). https://doi.org/10.1128/mbio.01781-15

16. Spinelli, S., Campanacci, V., Blangy, S., Moineau, S., Tegoni, M., Cambillau, C.: Modular Structure of the Receptor Binding Proteins of *Lactococcus lactis* Phages: The RBP Structure of the Temperate Phage TP901-1. *J. Biol. Chem.* 281, 14256–14262 (2006). https://doi.org/10.1074/jbc.M600666200
17. Hrebík, D., Stveráková, D., Füzik, T., Pantůček, P. Structure and genome ejection mechanism of Staphylococcus aureus phage P68. *Sci*. Adv.5,eaaw7414. (2019). https://doi.org/10.1126/sci-adv.aaw7414
18. Van Uffelen, A.: Discovering Phage Receptor-Binding Proteins in Metagenomics Data Using Machine Learning. Master's dissertation, Ghent University (2021). https://lib-store.ugent.be/fulltxt/RUG01/003/012/852/RUG01-003012852_2021_0001_AC.pdf
19. Klumpp, J., Dunne, M., Loessner, M. J. A perfect fit: Bacteriophage receptor-binding proteins for diagnostic and therapeutic applications. Curr. *Opin. Microbiol.*, 71, 1369-5274. (2023). https://doi.org/10.1016/j.mib.2022.102240.
20. Santos, S. B., Costa, A. R., Carvalho, C., Nóbrega, F. L., Azeredo, J. Exploiting bacteriophage proteomes: The hidden biotechnological potential. *Trends Biotechnol*., 36(9), 966-984. 2018. https://doi.org/10.1016/j.tibtech.2018.04.006.
21. Olawade, D. B., Fapohunda, O., Egbon, E., Ebiesuwa, O. A., Usman, S. O., Faronbi, A. O., Fidelis, S. C. Phage therapy: A targeted approach to overcoming antibiotic resistance. *Microb. Pathog*., 197, 0882-4010. (2024). https://doi.org/10.1016/j.micpath.2024.107088.
22. Harada, L. K., Silva, E. C., Campos, W. F., Del Fiol, F. S., Vila, M., Dąbrowska, K., Krylov, V. N., Balcão, V. M. Biotechnological applications of bacteriophages: State of the art. *Microbiol*. Res., 212–213, 38-58. (2018). https://doi.org/10.1016/j.micres.2018.04.007.
23. O'Sullivan, L., Buttimer, C., McAuliffe, O., Bolton, D., Coffey, A. Bacteriophage-based tools: recent advances and novel applications. *F1000Res*., 5, 2782. (2016). https://doi.org/10.12688/f1000research.9705.1.
24. Nováček, J., Šiborová, M., Benešík, M., Pantůček, R., Doškař, J., Plevka, P. Structure and genome release of Twort-like Myoviridae phage with a double-layered baseplate. *Proc. Natl. Acad. Sci. U.S.A.,* 113(33), 9351-9356. (2016). https://doi.org/10.1073/pnas.1605883113.
25. Ladunga, I. Finding similar nucleotide sequences using network BLAST searches. Curr. Protoc. *Bioinformatics*., 58, 3.3.1-3.3.25. (2017). https://doi.org/10.1002/cpbi.29.
26. Dorlass, E. G., Amgarten, D. E. Bioinformatic approaches for comparative analysis of viruses. *Methods Mol. Biol*., 2802, 395-425. (2024). https://doi.org/10.1007/978-1-0716-3838-5_13.
27. Prakash, A., Jeffryes, M., Bateman, A., Finn, R. D. The HMMER web server for protein sequence similarity search. *Curr. Protoc. Bioinformatics*., 60, 3.15.1-3.15.23. (2017). https://doi.org/10.1002/cpbi.40.
28. Pan, X., Shen, H. B. RNA-protein binding motifs mining with a new hybrid deep learning-based cross-domain knowledge integration approach. *BMC Bioinformatics*., 18, 136. (2017). https://doi.org/10.1186/s12859-017-1561-8.
29. Gonzales, M. E. M., Ureta, J. C., Shrestha, A. M. S. PHIStruct: improving phage–host interaction prediction at low sequence similarity settings using structure-aware protein embeddings. *Bioinformatics*, 41(1), btaf016. (2025). https://doi.org/10.1093/bioinformatics/btaf016.
30. Clark, J. R., March, J. B. Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials. *Trends Biotechnol*., 24(5), 212-218. (2006). https://doi.org/10.1016/j.tibtech.2006.03.003.
31. Hyman, P., Abedon, S. T. Bacteriophage host range and bacterial resistance. Adv. Appl. Microbiol., 70, 217-248. (2010). https://doi.org/10.1016/S0065-2164(10)70007-1.
32. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.
33. UniProt. https://www.uniprot.org/.
34. Broeker, N.K., Barbirz, S. Not a barrier but a key: how bacteriophages exploit host's O-antigen as an essential receptor to initiate infection. *Mol Microbiol* 105:353–357. (2017). https://doi:10.1111/mmi.13729

35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.,* 215(3), 403-410. (1990). https://doi.org/10.1016/S0022-2836(05)80360-2.

36. HMMER. http://hmmer.org/.

37. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci*, 27(1):135-145. (2017). https://doi:10.1002/pro.3290

38. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol., 2, 37-63. (2011).

39. Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589. (2021). https://doi.org/10.1038/s41586-021-03819-2.

40. Goulet, A., Spinelli, S., Mahony, J., Cambillau, C. Conserved and diverse traits of adhesion devices from Siphoviridae recognizing proteinaceous or saccharidic receptors. *Viruses* 12:512. (2020). https://doi:10.3390/v12050512

41. Zhang, Z., Introduction to Machine Learning: K-Nearest Neighbors. Annals of Translational Medicine, 4(11):218. (2016) https://doi.org/10.21037/atm.2016.03.37

42. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658-1659. (2006). doi:10.1093/bioinformatics/btl158

43. Breiman, L. Random forests. Mach. Learn., 45, 5-32. (2001). http://dx.doi.org/10.1023/A:1010933404324.

44. Cortes, C., Vapnik, V. Support-vector networks. Mach. Learn., 20, 273–297. (1995). https://doi.org/10.1007/BF00994018.

45. Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., (2016). https://doi.org/10.1145/2939672.2939785.

46. LeCun, Y., Bengio, Y. Convolutional networks for images, speech, and time series. Handb. Brain Theory Neural Netw., MIT Press, 255–258. (1998).

47. Hochreiter, S., Schmidhuber, J. Long short-term memory. Neural Comput., 9(8), 1735-1780. (1997). https://doi.org/10.1162/neco.1997.9.8.1735.

48. Degroux, S., Effantin, G., Linares, R., Schoehn, G., Breyton, C. Deciphering Bacteriophage T5 Host Recognition Mechanism and Infection Trigger. J Virol 97:e01584-22. (2023). https://doi.org/10.1128/jvi.01584-22

49. PDB. https://www.rcsb.org/.

50. Ackermann, H.W., "Tailed Bacteriophages: The Order Caudovirales," *Advances in Virus Research*, 51:135–201. (1998) https://doi.org/10.1016/S0065-3527(08)60785-X

51. Nobrega, F.L., Vlot, M., de Jonge, P.A. *et al.* Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol* 16, 760–773. (2018). https://doi.org/10.1038/s41579-018-0070-8

52. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, 305(3):567-580. (2001). https://doi:10.1006/jmbi.2000.4315

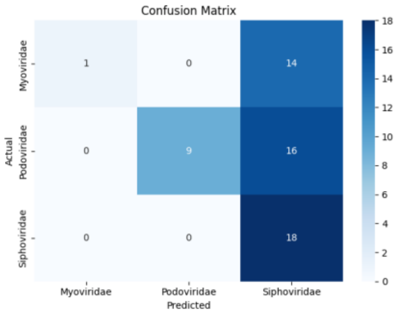
53. Ikotun, A., Ezugwu, A., Abualigah, L., Abuhaija, B., Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* 622, 178-210. (2023). https://doi.org/10.1016/j.ins.2022.11.139.

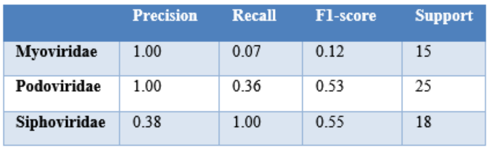
54. Sekhar, S.R.M., Siddesh, G.M., Raj, M. *et al.* Protein class prediction based on Count Vectorizer and long short term memory. *Int. j. inf. tecnol.* **13**, 341–348 (2021). https://doi.org/10.1007/s41870-020-00528-3

55. Hitchcock, N. M., Devequi Gomes Nunes, D., Shiach, J., Valeria Saraiva Hodel, K., Dantas Viana Barbosa, J., Alencar Pereira Rodrigues, L., Coler, B. S., Botelho Pereira Soares, M., & Badaró, R. (2023). Current Clinical Landscape and Global Potential of Bacteriophage Therapy. *Viruses*, *15*(4), 1020. https://doi.org/10.3390/v15041020

56. Taslem Mourosi, J., Awe, A., Guo, W., Batra, H., Ganesh, H., Wu, X., & Zhu, J. Understanding Bacteriophage Tail Fiber Interaction with Host Surface Receptor: The Key “Blueprint” for Reprogramming Phage Host Range. *International Journal of Molecular Sciences*, *23*(20), 12146. (2011). https://doi.org/10.3390/ijms232012146
57. Kamal, H., & Mashaly, M. Advanced Hybrid Transformer-CNN Deep Learning Model for Effective Intrusion Detection Systems with Class Imbalance Mitigation Using Resampling Techniques. *Future Internet*, *16*(12), 481. (2024). https://doi.org/10.3390/fi16120481
58. Allam, H., Davison, C., Kalota, F., Lazaros, E., & Hua, D. AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques. *Big Data and Cognitive Computing*, *9*(1), 16. (2025). https://doi.org/10.3390/bdcc9010016
59. Brüssow, H. Hurdles for Phage Therapy to Become a Reality—An Editorial Comment. *Viruses*, 11(6), p.557. (2019). https://doi.org/10.3390/v11060557
60. Gonzalez, F.; Scharf, B.E. Identification of Receptor Binding. Proteins in Flagellotropic Agrobacterium Phage 7-7-1. Viruses. 13, 1267. (2021). https://doi.org/10.3390/v13071267
61. Amaia Lasagabaster, Elisa Jiménez, Tatiana Lehnherr, Katherine Miranda-Cadena, Hansjörg Lehnherr. Bacteriophage biocontrol to fight Listeria outbreaks in seafood. Food and Chemical Toxicology. Volume 145. 111682. ISSN 0278-6915. (2020). https://doi.org/10.1016/j.fct.2020.111682.
62. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning [published correction appears in Sci Rep. (2024);14(1):15724. https://doi: 10.1038/s41598-024-66611-y.]. Sci Rep.;14(1):6086. (2024). https://doi:10.1038/s41598-024-56706-x
63. Groth D, Hartmann S, Klie S, Selbig J. Principal components analysis. Methods Mol Biol. 930:527-547. (2013). doi:10.1007/978-1-62703-059-5_22
64. Espinoza FA, Oliver JM, Wilson BS, Steinberg SL. Using hierarchical clustering and dendrograms to quantify the clustering of membrane proteins. Bull Math Biol. 74(1):190-211. (2012). https://doi:10.1007/s11538-011-9671-3

ANNEXES





	Precision	Recall	F1-score	Support
Myoviridae	1.00	0.07	0.12	15
Podoviridae	1.00	0.36	0.53	25
Siphoviridae	0.38	1.00	0.55	18

Fig. 1. Confusion Matrix and Classification Report. On the left we have a confusion matrix that shows classification performance of the model across the three phage morphotypes: Myoviridae, Podoviridae, and Siphoviridae. True labels are shown in the rows and predicted labels are shown in columns. Correct classifications appear along the diagonal, and misclassifications are shown off-diagonal. Colour intensity reflects the number of instances, with darker shades indicating higher values. The colour bar on the right provides the numerical scale. The classification report (right) summarizes performance of the model by using Precision (right positive predictions proportion), Recall (real positives correctly identified proportion), F1-score (harmonic mean between recall and precision), and Support (true instances for each class) [58,62].