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 mediate bacterial recognition 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. RBP-ID aims to integrate multiple computational approaches, including DIAMOND for homology searches, unsupervised machine learning like K-means for RBP clustering and supervised machine learning like KNN to measure similarity. By leveraging validated and predicted RBP datasets, it will 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 State of the Art

Bacteriophages, or phages, are viruses that specifically infect bacterial hosts, playing a critical role in regulating bacterial populations and shaping microbial community dynamics in environments such as marine ecosystems, soil, and the human microbiome [1-3,9,11]. 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 between phages and their bacterial hosts [11,18]. 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].

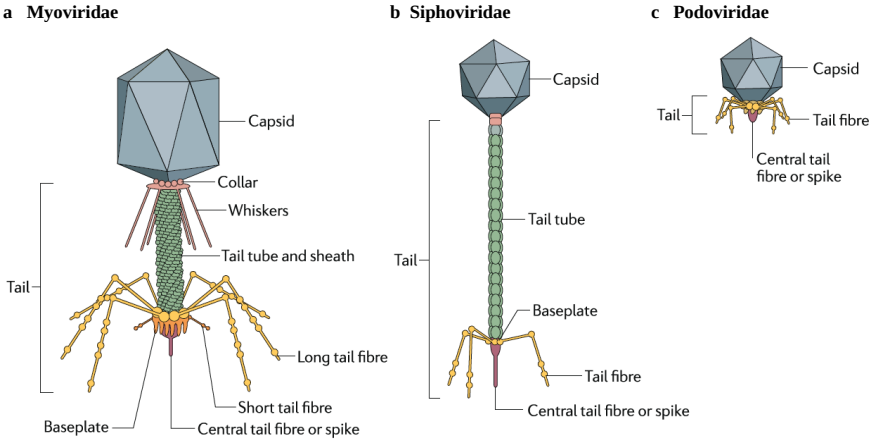

Fig. 1. Structures of the three tailed morphotypes of the order *Caudovirales*. (a) Myoviridae is the only family with a contractile tail. (b) Siphoviridae have a baseplate at the distal end of the tail, similarly to Myoviridae, yet has a non-contractile tail. (c) Podoviridae is the only family without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

The structural variation in RBPs contributes to the specificity of different phage morphotypes, such as Myoviridae**,** Siphoviridae, and Podoviridae, all members of the classe of tailed viruses *Caudoviricetes*. Myoviridae is characterized by contractile tail fibers and a complex baseplate which employes 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, receptor-binding proteins (RBPs) are key targets for engineering phages with altered host ranges, enabling the development of synthetic phages 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].

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].

Machine learning classifiers, such as Random Forest and Support Vector Machines (SVMs), have been employed 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, two distinct datasets are being compiled separately: one consisting of experimentally validated phage RBPs and another containing computationally predicted phage RBPs.

2.1.1 Validated and Computed RBP Dataset

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

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

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

- **NCBI**: A broad search for RBPs will be 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" will be used to find annotated RBP proteins. Filters ensured relevance to Caudoviricetes and only experimentally validated structures will be 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 will include 13 variables defined a priori namely: data_origin, phage_accession_nr, phage_name, rbp_name, rbp_accession_nr, protein_type, tax_class_phage, morphotype, genome_id_phage, phage_sequence (nr), rbp_sequence (aa), delineation (aa), annotation, host.

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

2.1.2 Homology-Based Identification and Clustering

This phase of the project focuses on analysing sequence similarity using DIAMOND, a tool for homology-based identification of receptor-binding proteins (RBPs). This method compares validated and predicted RBPs to determine similarity scores, helping to identify evolutionary relationships and potential homologous proteins [42].

Secondly, K-means clustering, an unsupervised machine learning algorithm, will be used to classify RBPs based on their similarity [54]. The goal is to group similar proteins while minimizing distance between data points and cluster centroids. Additionally, k-Nearest Neighbours (KNN), a supervised machine learning method, may be applied to measure similarity using the Euclidean Distance metric. This will help identify distinct subgroups among tail fibers based on sequence and structural features [41].

To facilitate computational analysis, a FASTA file containing RBP sequences will be created. Statistical evaluations, including alignment-based p-value assessments, will validate the accuracy of clustering and sequence similarity [38].

By combining sequence similarity analysis, machine learning clustering, and structural validation, this methodology aims to improve RBP identification, classification, and functional annotation in bacteriophages.

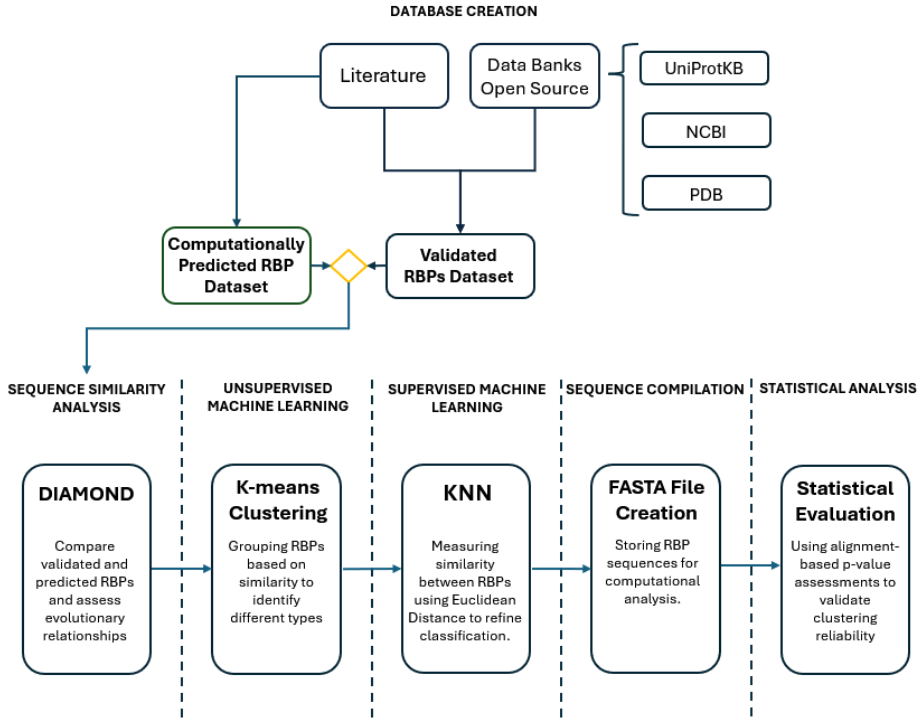

Fig. 2. Workflow for RNA-Binding Protein Dataset Construction. This workflow integrates literature reviews, public databases (UniProtKB, NCBI, PDB), and computational predictions to create RNA-binding protein (RBP) datasets. Sequence similarity is assessed using DIAMOND, followed by clustering (K-means) and supervised classification (KNN). Data is compiled into FASTA format, and statistical validation ensures reliability in the final dataset.

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/sciadv.aaw7414
18. Van Uffelen, A.: Discovering Phage Receptor-Binding Proteins in Metagenomics Data Using Machine Learning. Master's dissertation, Ghent University (2021). https://libstore.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. Buchfink, B., Reuter, K. & Drost, HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18, 366–368 (2021). https://doi.org/10.1038/s41592-021-01101-x
43. Rigsby, R.E., Parker, A.B. Using the PyMOL application to reinforce visual understanding of protein structure. *Biochem Mol Biol Educ*. ;44(5):433-437. (2016). https://doi:10.1002/bmb.20966
44. Breiman, L. Random forests. Mach. Learn., 45, 5-32. (2001). http://dx.doi.org/10.1023/A:1010933404324.
45. Cortes, C., Vapnik, V. Support-vector networks. Mach. Learn., 20, 273–297. (1995). https://doi.org/10.1007/BF00994018.
46. 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.
47. LeCun, Y., Bengio, Y. Convolutional networks for images, speech, and time series. Handb. Brain Theory Neural Netw., MIT Press, 255–258. (1998).
48. 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.
49. 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
50. PDB. https://www.rcsb.org/.

51. 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
52. 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
53. 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
54. 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.