PhageScan - Automated viral particle detection

Sousa, R.[1][0009−0000−2558−7176], Dias, O.[1][0000−0002−1765−7178], and Oliveira, H.[1][0000−0001−9564−5714]

Department of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal sac@ceb.uminho.pt
https://www.deb.uminho.pt/

Abstract. Bacteriophages emerged as a promising solution for the looming threat of antibiotic-resistant bacteria. Being the most abundant organism in the biosphere, classifying these viruses is a task of relative complexity for which computational tools are playing a pivotal role, aiding the understanding of the complexities of phage biology and taxonomy. This work proposes the implementation of YOLO algorithms for automatic phage identification from SEM / TEM images, resulting in a user-friendly tool to obviate the laboriousness and time-consumption of manual classification. In its first iteration, the tool will focus on the identification of siphoviruses, given the morphotype's abundance and morphological consistency, expecting to classify siphoviruses on micrographs as well as perform some basic measurements of capsid size. This article describes the approach utilized in this development, from data set construction to the implementation of the algorithms.

Keywords: Bacteriophage classification · deep learning · convolutional neural networks

1 Background

1.1 What are bacteriophages?

Bacteriophages (phages), regarded the most abundant organisms in the biosphere, are a type of virus that thrives by infecting and replicating within bacterial cells. [1] These organisms are species-specific, typically targeting a single bacterial species to maintain their lifecycle and co-evolving with it. [2,3,4,5]

Phages adopt two distinct life styles: lytic or lysogenic (Fig 1 [6]). Lytic phages infect the host cell and hijack its biosynthetic machinery to rapidly generate offspring that exit the cell, bursting and killing it in the process. Phages with a lysogenic life style, by their turn, can be stable for generations, integrating their genome in the host's, sometimes even shaping the host phenotype. In this state, they do not necessarily kill the host, adopting an opportunistic approach, transitioning towards a lytic approach when the host reaches its end of life and / or if a specific trigger occurs. There are also phages that exhibit a a chronic life style, which is generally non-bactericidal. [2,6,7]

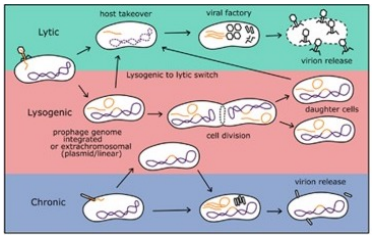


Fig. 1: Simple schematic representation of the different phage life cycles. [6]

Given their natural predating relationship with bacteria, phages are increasingly gaining traction as potential therapeutic agents in the combat of antibiotic-resistant bacteria, a pressing global health challenge.[7,8,9,10] A result of this growing interest in phages creates a pressing need to develop methods to study phages to facilitate identification and classification.

1.2 Phage classification

With the advancements in genomics and metavirunomics, allowing for a more accurate and comprehensive understanding of phage diversity, phage classification has undergone significant changes in recent years, transitioning from a morphological approach to a genomically coherent taxonomical method. The biggest change in the system was the abolishment of former family of tailed phages, caudovirales (Fig 2 [11]), the most abundant group of phages, which was divided into three subfamilies: (i) podoviridae; (ii) myoviridae; and (iii) siphoviridae. In its stead emerged the new class of caudiviricetes, comprised of 4 new orders and 22 new families. [11,12,13,14]

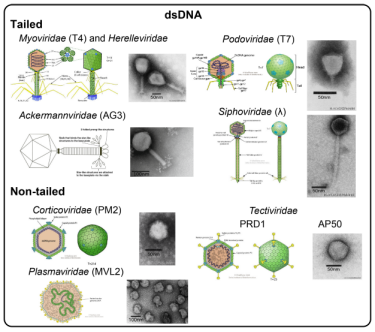
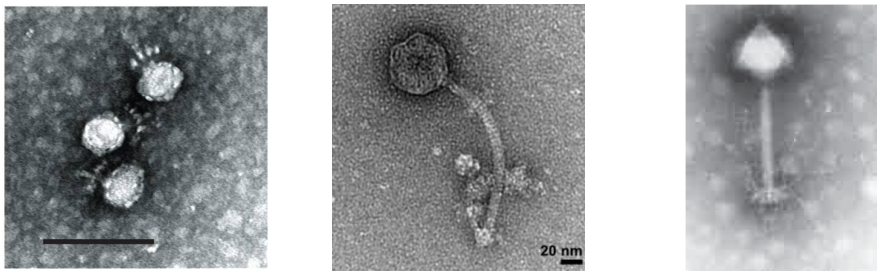


Fig. 2: Schematic representation of tailed phages morphology. [11]

While morphology is no longer used in phage taxonomy, it still retains its importance and impact towards identification and classification and standout as a potential approach for automating tailed phage classification from TEM images. For this work we will use a morphological approach to identify phages based on their morphotype, aiming to identify phage morphotype from TEM images and perform approximate measurements to capsid volume and genome size. There are three distinct tailed phage morphotypes (Fig 3), based on their respective tailed structures: podoviruses, siphoviruses and myoviruses. [15] **Podoviruses** typically display a short non-contractile tail and an icosahedral capsid of about 60 nm in diameter. An example of this virus is *Brucealvirus* CP7R as shown below. [16] **Siphoviruses** typically display an icosahedral capsid and a long, flexible but non-contractile tail, spiked at the tip. An example if this morphotype, BlueFeather, can be seen below. [17] **Myoviruses** display a contractile tail with a base plate and spikes at the tip and an icosahedral capsid. One example of this viruses is *Synechococcus* Phage S-PM2. [18]



(a) Podovirus (b) Siphovirus (c) Myovirus

Fig. 3: Morphotypes of phages present in the caudoviricetes class, the most abundant class of phages in the currently existing knowledge. From left to right are shown specimens of podovirus CP7R, a Chlostridium perfringens phage, siphovirus BlueFeather, a Arthrobacter globiformis phage, and myovirus S-PM2, a Synecochus sp. phage. [16,17,18]

As for viral particle size, specifically capsid sizes, these can range from 30 nm to 180nm, enveloping genomes of varying sizes and complexities. [19,20] While there is no clear consensus about the direct correlation between genome and capsid sizes, several authors have been registering a generally accepted allometric relationship between genome length and capsid volume. For the sake of this study we will consider a ratio of 1.52 capsid volume: genome length, being this the adjusted exponent found for dsDNA viruses, specifically tailed phages. [19,20,21,22]

1.3 Automated viral particle detection and classification

Automatic detection and recognition of particles from digital microscopy images is still an active yet unexplored field. One of the main advantages of developing accurate and efficient methods for automatic detection of viral particles is the automation of the fastidious and labour-intensive work associated with microscope analysis - safeguarding the impact that human factors such as fatigue can have on the analysis (be it counting or identifying particle types). [23] This is also the case for automatic recognition of viral particles, where qualitative and quantitative measurements involve processing hundreds or thousands of microscopy images to count particles and identify viral morphotypes. [23,24]

It is accepted that in order to be regarded as valuable, automatic recognition systems must attain an accuracy greater than 75% when compared to a human expert. [23] There are notable methods in the literature for automatic particle identification using machine learning techniques such as neural networks, Gaussian classifiers, Naive Bayes, and random forests that have achieved this requisite to different degrees. [24,25,26,27]

Each of these techniques exhibit different advantages and disadvantages considering the scenarios and data at hand, for instance: Gaussian classifiers, operating as statistical models, excel at classifying image features assuming a Gaussian distribution. These classifiers assess the probability of each data point belonging to a specific class based on its features and assign it to the most probable class. They are particularly effective when handling images with features that adhere to a normal distribution and classes that are distinctly separated. [28]

Random forests, on the other hand, are ensemble learning approaches that can be used for classification and regression. Each decision tree in the ensemble is trained on a random subset of the training data and features. During prediction, the forest's trees separately anticipate the class of the incoming data, with the final prediction established by a majority vote or averaging across all trees. Random forests, which are well-known for their resistance to over fitting, are especially useful when dealing with heterogeneous datasets. [29]

Naive Bayes, despite its simplicity, emerges as a powerful probabilistic classifier. It operates under the assumption of feature independence, calculating class probabilities based on the input image data and selecting the class with the highest probability as the prediction. While straightforward, Naive Bayes demonstrates remarkable proficiency, especially in scenarios involving image classification tasks. [30]

Finally, neural networks, specifically convolutional neural networks (CNNs) inspired by the human brain, stand as versatile models for image analysis. Through interconnected layers of nodes, neural networks learn complex patterns and relationships within images by adjusting connection weights during training. Their adaptability and capability to discern intricate patterns have propelled their success across various domains, including image recognition and particle identification. [31]

2 Methods/approach

This work aims to explore Machine Learning (ML) and Deep Learning (DL) techniques and develop a novel tool - PhageScan - to classify phages, more specifically, we to evaluate the employment of CNNs in automated phage particle recognition given their versatility and high level of discernment for image classification with high accuracy. [25] More specifically we will explore and evaluate the application of You Only Look Once (YOLO) algorithms. YOLO was created as an approach for object detection and has been demonstrating growing accuracy and real-time performance and have already been employed in the field of biological images processing. [32,33,34]

It is important to refer the existence of several challenges for the development of this tool namely: (i) the lack of publicly available data sets of phage TEM micrographs, the existing phage micrographs are sparse and spread throughout data bases and / or closed access publications, thus raising copyright issues; (ii) the quality of available TEM micrographs, which tipically have high levels of noise and require severe processing; (iii) the lack of a good variety of micrographs of phages in their environment, namely multiple phages and hosts on a sample; (iv) podoviruses exhibit small tails that could be confused be hard to detect; (v) myoviruses' tails have a contracted / relaxed state, adding complexity to labelling and classification. On the other hand, siphoviruses, with their relatively consistant capsid shape and long non-contractile tails, are a good starting point, moreover siphoviruses are the most abundant morphotype.

2.1 Data gathering, preprocessing and labelling

Data gathering The success of training an object-detection algorithm such as YOLOv8 lies deeply in the amount and quality of the data set that is presented to the model. As such, it is crucial to have an extensive and diverse range of images that adequately represent the variability and complexity of scenarios in which the tool will be used. Specifically, it is fundamental to gather large amounts of TEM micrographs of isolated siphoviruses, siphoviruses and bacteria (infection scenario for example), siphoviruses and other phage types, among other scenarios. Having this diversity and consistency ensures better contextual identification of myoviruses, reduces false positives and ensures better accuracy. This in itself is a challenging undertaking as TEM micrographs with these specific scenarios are relatively scarce. As such, on a first approach we will be sourcing TEM micrographs from curated databases such as PhagesDB which has a repertoire of 243 phages with TEM images 119 of which belonging to the siphovirus morphotype. Existing literature studying specific myoviruses can also be a source.

2.2 Preprocessing, augmentation and annotation

Once the data set creation is achieved, the next step is to preprocess the data, ensuring the consistency and quality of the images contained by the data set

through resizing images to a fixed size, ensuring all scales are correctly represented, adjusting image quality (brightness, contrast, noise and artifact removal). The data set will also be augmented through random rotations and flips in order to create scenarios where the target objects of classification and identification can be identified independently from image rotation angle, being this a common data augmentation approach for object detection algorithms. [34]

Finally, to make the data set presentable to the algorithm, it's fundamental to annotate the images in order to specify what is what. This process is achieved through specifying object classes and location within the images with bounding boxes. This can be achieved using tools such as using tools such as LabelImg or Roboflow, being the latter well documented in the context of data set preparation and annotation for YOLO algorithms. The annotation process for phages can follow different approaches: (i) draw a bounding box directly on top of the phage, the simplest approach to classify phages in micrographs where there are other objects present; (ii) separating phages into two sub-classes: head and tail, particularly relevant for images of isolated phages, increasing context in micrographs where could exist other objects with strong contrast resembling the head such as bacteria. Considering we will be dealing with TEM micrographs, cases in which there are images with high amounts of noise and incorrect contrast levels could arise. This would imply considering the application of more complex image processing techniques typically used in computational analysis of microscopy images, such as edge detection/filtering, pixel intensity interpolation. These techniques refine the image in order to make the phage objects (head and tail) more visible. [26] It is important to reserve smaller, unannotated subsets of the dataset to serve as validation and testing sets, respectively. This allows for an accurate assessment of the algorithm's effectiveness in training, validation, and generalization to unseen data.

2.3 Model training

Two object detection architectures will be tested: YOLOv5 and YOLOv8. The models will then be trained on the generated data set for a varying number of epochs in order to empirically determine the best training settings and compare results for both algorithms in order to find the most efficient for the desired application. For this comparison, we will be evaluating and comparing the models' performance, namely by analysing each model's accuracy, precision, recall, F1-Scores and confusion matrices. Once the classification component of the tool is working with a satisfactory accuracy we will then move on to the implementation of automatic capsid annotation and measurement for genome size estimation.

References

1. S. Batinovic, F. Wassef, S.A. Knowler, D.T.F. Rice, C.R. Stanton, J. Rose, J. Tucci, T. Nittami, A. Vinh, G.R. Drummond, C.G. Sobey, H.T. Chan, R.J. Seviour, S. Petrovski, and A.E. Franks. Bacteriophages in natural and artificial environments. *Pathogens*, 8(3):100, 2019.
2. M.R. Clokie, A.D. Millard, A.V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, 2011.
3. L.M. Kasman and L.D. Porter. Bacteriophages. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-., 2022. Available from: url-https://www.ncbi.nlm.nih.gov/books/NBK493185/.
4. B. Koskella and M.A. Brockhurst. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev*, 38(5):916–931, 2014.
5. C. Suttle. Viruses in the sea. *Nature*, 437:356–361, 2005.
6. C. Howard-Varona, K.R. Hargreaves, S.T. Abedon, and M.B. Sullivan. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J*, 11(7):1511–1520, 2017.
7. P. Ioannou, S. Baliou, and G. Samonis. Bacteriophages in infectious diseases and beyond—a narrative review. *Antibiotics*, 12:1012, 2023.
8. I.U. Haq, W.N. Chaudhry, and M.N. et al. Akhtar. Bacteriophages and their implications on future biotechnology: a review. *Virol J*, 9:9, 2012.
9. S.B. Gamachu and M. Debalo. Review of bacteriophage and its applications. *Int J Vet Sci Res*, 8(3):133–147, 2022.
10. A. Sulakvelidze, Z. Alavidze, and J.G. Jr. Morris. Bacteriophage therapy. *Antimicrob Agents Chemother*, 45(3):649–659, 2001.
11. M.B. Dion, F. Oechslin, and S. Moineau. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol*, 18(3):125–138, 2020. Epub 2020 Feb 3.
12. D. Turner, A.M. Kropinski, and E.M. Adriaenssens. A roadmap for genome-based phage taxonomy. *Viruses*, 13(3):506, 2021.
13. D. Turner, A.N. Shkoporov, and C. et al. Lood. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ictv bacterial viruses subcommittee. *Arch Virol*, 168:74, 2023.
14. H.W. Ackermann. 5500 phages examined in the electron microscope. *Arch Virol*, 152(2):227–243, 2007.
15. E. Maffei, A. Shaidullina, M. Burkolter, V. Druelle, L. Willi, F. Estermann, S. Michaelis, H. Hilbi, D. Thaler, and A. Harms. Systematic exploration of escherichia coli phage-host interactions with the basel phage collection. 2021. doi: 10.1101/2021.03.08.434280.
16. N.V. Volozhantsev, B.B. Oakley, C.A. Morales, V.V. Verevkin, V.A. Bannov, V.M. Krasilnikova, A.V. Popova, E.L. Zhilenkov, J.K. Garrish, K.M. Schegg, R. Woolsey, D.R. Quilici, J.E. Line, K.L. Hiett, G.R. Siragusa, E.A. Svetoch, and B.S. Seal. Molecular characterization of podoviral bacteriophages virulent for clostridium perfringens and their comparison with members of the picovirinae. *PLoS One*, 7(5):e38283, 2012.
17. S. Demo, A. Kapinos, A. Bernardino, K. Guardino, B. Hobbs, K. Hoh, E. Lee, I. Vuong, K. Reddi, A. Freise, and J. Parker. Bluefeather, the singleton that wasn't: Shared gene content analysis supports expansion of arthrobacter phage cluster fe. *PLOS ONE*, 16:e0248418, 2021.
18. N. Mann. The third age of phage. *PLoS biology*, 3:e182, 2005.

19. A. Luque, S. Benler, D.Y. Lee, C. Brown, and S. White. The missing tailed phages: Prediction of small capsid candidates. *Microorganisms*, 8:1944, 2020.
20. D.Y. Lee, C. Bartels, K. McNair, R.A. Edwards, M.A. Swairjo, and A. Luque. Predicting the capsid architecture of phages from metagenomic data. *Comput Struct Biotechnol J*, 20:721–732, 2022.
21. J. Cui, T.E. Schlub, and E.C. Holmes. An allometric relationship between the genome length and virion volume of viruses. *J Virol*, 88, 2014.
22. H.V. Chaudhari, M.M. Inamdar, and K. Kondabagil. Scaling relation between genome length and particle size of viruses provides insights into viral life history. *iScience*, 24(5):102452, 2021.
23. R.M. Glaeser. Historical background: why is it important to improve automated particle selection methods? *Journal of Structural Biology*, 145(1–2):15–18, 2004.
24. A. Gelzinis, A. Verikas, E. Vaiciukynas, M. Bacauskiene, S. Sulcius, E. Simoliunas, J. Staniulis, and R. Paskauskas. Automatic detection and morphological delineation of bacteriophages in electron microscopy images. *Computers in Biology and Medicine*, 64:101–116, 2015.
25. Toshihiko Ogura and Chikara Sato. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis. *Journal of Structural Biology*, 145(1–2):63–75, 2004.
26. C.O.S. Sorzano, E. Recarte, M. Alcorlo, J.R. Bilbao-Castro, C. San-Martín, R. Marabini, and J.M. Carazo. Automatic particle selection from electron micrographs using machine learning techniques. *Journal of Structural Biology*, 167(3):252–260, 2009.
27. César A.B. Castañón, Jane S. Fraga, Sandra Fernandez, Arthur Gruber, and Luciano da F. Costa. Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus eimeria. *Pattern Recognition*, 40(7):1899–1910, 2007.
28. Lixiang Xu, Biao Zhou, Xinlu Li, Zhize Wu, Yan Chen, Xiaofeng Wang, and Yuanyan Tang. Gaussian process image classification based on multi-layer convolution kernel function. *Neurocomputing*, 480:99–109, 2022.
29. Peter D. Caie, Neofytos Dimitriou, and Ognjen Arandjelović. Precision medicine in digital pathology via image analysis and machine learning. In Stanley Cohen, editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 149–173. Elsevier, 2021.
30. Gilberto Francisco Martha de Souza, Adherbal Caminada Netto, Arthur Henrique de Andrade Melani, Miguel Angelo de Carvalho Michalski, and Renan Favarão da Silva. Chapter 6 - engineering systems' fault diagnosis methods. In Gilberto Francisco Martha de Souza, Adherbal Caminada Netto, Arthur Henrique de Andrade Melani, Miguel Angelo de Carvalho Michalski, and Renan Favarão da Silva, editors, *Reliability Analysis and Asset Management of Engineering Systems*, Advances in Reliability Science, pages 165–187. Elsevier, 2022.
31. Jonas Teuwen and Nikita Moriakov. Chapter 20 - convolutional neural networks. In S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger, editors, *Handbook of Medical Image Computing and Computer Assisted Intervention*, The Elsevier and MICCAI Society Book Series, pages 481–501. Academic Press, 2020.
32. Xiyao Li, Jingwen Chen, Yong He, Guofeng Yang, Zhongren Li, Yimin Tao, Yanda Li, Yu Li, Li Huang, and Xuping Feng. High-through counting of chinese cabbage trichomes based on deep learning and trinocular stereo microscope. *Computers and Electronics in Agriculture*, 212:108134, 2023.

33. R. Zhu, Y. Cui, J. Huang, E. Hou, J. Zhao, Z. Zhou, and H. Li. Yolov5s-sa: Light-weighted and improved yolov5s for sperm detection. *Diagnostics (Basel)*, 13(6):1100, 2023.
34. D.G. Gonzalez et al. Evaluating rotation invariant strategies for mitosis detection through yolo algorithms. In A. Cunha, M. Garcia, N. Marx Gómez, and S. Pereira, editors, *Wireless Mobile Communication and Healthcare*, volume 484 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Cham, 2023. Springer.