PhageLysin: Phage Endolysin Finder

Tiago Baptista[1], Hugo Oliveira[2], and Óscar Dias[2]

[1] University of Minho, Portugal
[2] University of Minho, Center of Biological Engineering (CEB), Portugal

1 Introduction

1.1 Aims

Bacterial multidrug resistance to antibiotics is one of the most concerning problems faced in modern days, however new alternative therapies are being developed. The potential use of enzybiotics, a relatively new term that defines therapies using phage-encoded enzymes, normally phage lytic enzymes, is rising as arguably the best alternative to regular antibiotics. One of those enzymes is the endolysin, utilized by phages to degrade the cell wall from within. There are also several studies that have shown that recombinant endolysins can also be applied to degrade cell wall from the outside, thus confirming their antibacterial properties [1] [2].

This project aims at building a comprehensive database of phage endolysins. The ultimate goal, probably outside of the scope of this project, would be the development of a machine learning (ML) tool to provide a quick and reliable way to identify endolysins in new phage genomes, but also on metagenomic sequencing data. The main steps of this project include: (i) review of public datasets containing metagenomic data; (ii) review of annotation tools applied to phage genomes on metagenomic data; (iii) build a database of phage endolysins; (iv) explore ML approaches to improve the accuracy of predictions i.e. classify endolysins.

1.2 Phages and endolysins

Phages represent the most plentiful and varied biological entities on the planet and are thought to exist in virtually every imaginable environment. It is estimated that there are 10^{31} phages, outnumbering their bacterial hosts by an estimated tenfold [9] [20] [25]. Phages exhibit remarkable diversity with a wide range of genome sizes from as low as 2,435 bp to as high as >540 kb [9].

Phages classification is based on various characters that describe the virus and allow to differentiate one from another. Characters include the molecular composition of the genome; the structure of the virus capsid and whether or not it is enveloped; the gene expression program used to produce virus proteins; host range; pathogenicity; and sequence similarity. Even though all characters are important in determining taxonomy, sequence comparisons using both pairwise

sequence similarity and phylogenetic relationships are the main sets of characters to define and distinguish virus taxa [17]. Furthermore, phages taxonomy is a complex subject and is regularly under revision by the International Committee on Taxonomy of Viruses (ICTV) [28]. A recent ICTV taxonomy update constituted the class *Caudoviricetes* as one of the most important ones, encompassing all tailed bacterial and archaeal viruses with icosahedral capsids and a doublestranded DNA (dsDNA) genome [26].

Phages replication first requires infection of the bacteria so they must bind to the bacterium surface and subsequently resort to a spike shaped complex of proteins at the bottom of the baseplate of phage's structure that punctures into the bacterial host, allowing a release of phage's genetic material into the host's intracellular environment. After that, phages can undergo one of two possible life cycles: lytic or lysogenic. The lytic cycle is followed by strictly lytic phages (also termed virulent phages) and consists in the seizure of bacterial host's metabolic machinery, which is first redirected to replication of phage's genome to create multiple copies of its own genetic material. Then, translation occurs and the synthesis of viral proteins and enzymes is achieved. Among them, some are used to assemble capsid and tails of new phages and package their genome, while enzymes such as holins and endolysins allow bacterium lysis and newly assembled phages release. In turn, the lysogenic cycle exhibited particularly by temperate phages requires the integration of the viral genome within the host's genome. This lysogenic bacteria has an integrated prophage within its genome, that can remain dormant through numerous bacterial cell divisions, unless triggered by stress or cellular damage in the bacterial host. Upon activation by these factors, the phage cycle shifts towards replication through a lytic pathway, after excision from the bacterial genome, culminating in the release of the new phage particles [13] [25].

It is thought that the holin-endolysin lytic system, also known as the lambda paradigm is universal in almost all dsDNA phages. As depicted in Fig. 1, this system consists of endolysin accumulation in the intracellular space due to the bacterial inner membrane. This accumulation goes on until the holin, a small hydrophobic membrane-spanning protein, after being expressed at a genetically programmed time, had accumulated in clusters producing homo-oligomeric pores in the inner membrane. Holin hole formation triggers the activation of the endolysin by granting its access to the peptidoglycan (PG) (or murein), a heteropolymer consisting of disaccharidepeptide repeat units linked by glycosidic bonds that form glycan strands. These strands are cross-linked through pentapeptides by 4–3 and 3–3 linkages. Endolysin muralytic activity i.e. its capability to degrade the PG, results in a unsustainable internal osmotic pressure, leading to cell lysis and progeny release. [4] [20]. Endolysins classification based on their lytic activity can be categorised into glycosidases, amidases and endopeptidases [1] [20] as shown in Fig. 1.

Endolysins unique ability to rapidly kill bacteria in a species specific manner puts them as promising antibacterial and biocontrol agents with applications in fermentations, food preservation, biotechnology, and medicine. Studies

regarding endolysin capacity to lyse bacteria when externally added were already performed. In the case of Gram-positive bacteria, cell lysis was already achieved *in vitro*, leading to the complete death of a streptococcus culture in a few seconds [20]. Other *in vivo* studies have shown endolysins efficacy against a variety of Gram-positive bacteria such as PlyC, C1, ClyR, Cpl-1, ClyV, and ClyJ [1]. Regarding Gram-negative bacteria, exogenous action of endolysins is still restricted due to the presence of the impermeable OM, constituting one of the most important challenges in endolysin therapy, however in recent years molecular engineering approaches have increased the applicability of endolysins in targeting Gram-negative bacteria [1] [20].

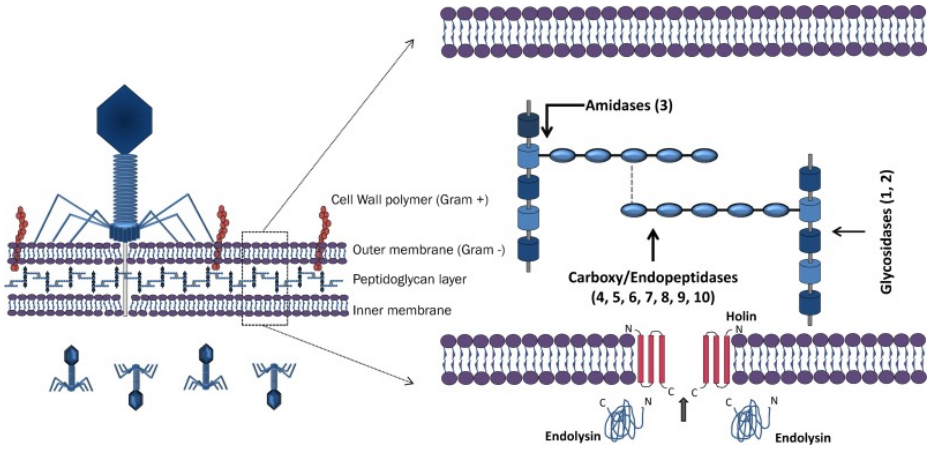

Fig. 1. Schematic representation of phage endolysins access to the PG through the holin-endolysin lytic system, alongside with a generic PG structure illustration of endolysins cleavage sites [20].

1.3 Phages and metagenomics

The nonexistence of a conserved genetic marker combined with the predicted large number of existing phages makes phage genomic diversity really difficult to comprehend [9]. Viral metagenomics emerged as a means to assess phage diversity and indirectly abundance [6], overcoming culture-based approaches and single marker genes challenges [9]. This approach consists of collecting, concentrating and sequencing the total viral component/population from a specific environment [6] [14]. This became possible due to optimization of required steps to obtain good quality viral nucleic acids, progress in sequencing technologies combined with more affordable costs and the rise and improvement of a set of analytical tools, allowing the construction of large-scale viral datasets of viral communities [9].

Metagenomics can be used to identify phages sequences and phage genes with higher or lower expression in a specific environment, which provides a good entry point to understand phages roles and reducing the need to culture phage hosts and isolate phages. It can also provide information on phages not prone to propagation, or phages with hosts that cannot be cultured in the lab [6].

2 Materials and Methods

2.1 Phage endolysins database

We expect to build a solid endolysins database that can be used to train the machine learning models, by querying existing databases containing endolysins.

Phage lytic proteins databases containing endolysins are available, namely phiBIOTICS [15], EnzyBase (renamed to EnzyBase2) [30], GMEnzy [29], PhaLP [8] and PhalydDB [12]. phiBIOTICS, EnzyBase2 and GMEnzy data is collected manually so they naturally have few entries. Oppositely, PhaLP and PhalydDB are comprehensive databases with PhaLP being automatically updated with each new UniProt release, thus containing a significantly bigger amount of entries [8] [12]. It would be possible to construct the endolysins database by querying the previously mentioned databases. PhaLP allows database querying on MySQL [8]. phiBIOTICS, EnzyBase2 and GMEnzy are not optimised for that so manual retrieving of information, use of APIs or web scraping could be possible ways to query these databases. PhalydDB is still a recent project with [12] having a data sharing statement indicating the necessity to request their authors access to their data.

Even though it is not a specialized database for phage lytic proteins, endolysins can be queried from National Center for Biotechnology Information (NCBI)'s Protein database, for example using the Biopython package, thus providing an alternative way to build or complete the endolysins database.

2.2 Metagenomic datasets

With the rise of culture independent approaches, a great quantity of phage metagenomic datasets has emerged. In the scope of this project, we will review some of them that constitute potential usable datasets for our work.

In a study by Fernandez-Ruiz and collaborators [10] a database of uncultured viral genomes was compiled from various sources. The dataset comprised 183,298 genomic sequences of uncultured viral genomes and they also compiled metadata associated with those sequences. Nayfach and collaborators [19] performed a large-scale identification of viral genomes from 11,810 bulk metagenomes from human stool samples obtained from multiple studies. With these data they built the Metagenomic Gut Virus (MGV) catalogue gathering an astonishing 189,680 viral draft genomes. In this work of Camarillo-Guerrero and collaborators [5] they introduce the Gut Phage Database, with approximately 142,000 non-redundant viral genomes gathered from a dataset of 28,060 global human gut metagenomes

and 2,898 reference genomes of cultured gut bacteria. Roux and collaborators [24] mined the viral signal from 14,977 publicly available bacterial and archaeal genomic datasets using VirSorter and yielding 12,498 high-confidence viral sequences with known hosts. Paez-Espino and collaborators [21] analysed metagenomic sequence data from 3,042 geographically diverse samples uncovering over 125,000 partial DNA viral genomes. Coutinho and collaborators [7] reported a dataset of 27,346 marine virome contigs that included 44 complete genomes of phages from marine environments.

2.3 Annotation tools applied to phages genomes on metagenomic data

Fernandez-Ruiz and collaborators [10] were able to identify putative endolysins using a database of uncultured viral genomes, alongside with related available metadata and a dataset of endolysin sequences. They used Prodigal [16] on metagenomic mode to identify protein encoding genes of uncultured phage genomes. Then, queried predicted protein sequences against the reference endolysin database using Diamond [3]. To filter results and determine putative endolysins, they applied the following thresholds: identity $x \geq y$ 50%, e-value $x \leq y$ 0.001, query coverage $x \geq y$ 30%, and alignment length $x \geq y$ 50 amino acids. They queried sequences against the Pfam database using HMMER version v3.1b2 [11] with default parameters to identify protein domains and used SignalIP [23] to detect signal peptide sequences. This work could be of great importance to our project, because it provides a replicable methodology to identify putative endolysins in phage/viral genomes, by using our future built endolysins database and a selected metagenomic dataset with phage genomes.

Despite that, it would also be necessary a way to identify phage genomes in metagenomic data. Luckily, there are a great variety of tools to identify phages on metagenomic sequencing data. Ho and collaborators performed a comparative analysis on 10 tools with such purpose, namely DeepVirFinder, Kraken2, MetaPhinder, PPR Meta, Seeker, ViralVerify, VIBRANT, VirFinder, VirSorter and VirSorter2. Results suggested that Kraken2 should be considered to identify previously characterised phages, while for novel phage detection Kraken2 should be used in combination with tools such as VirSorter2 and DeepVirFinder.

Furthermore, we are still missing phage genome annotation tools to retrieve features to use on the ML models. MetaPhage [22] is a fully automated computational pipeline for phage detection, classification, and quantification of metagenomics data. It was designed for scalability and reproducibility, while it is modular, allowing users to only perform certain chunks of the pipeline. These characteristics are very interesting, because they could allow the usage of this pipeline for annotation purposes only. On top of that, it incorporates previously referred tools.

2.4 Machine learning approaches to improve the accuracy of predictions i.e. classify endolysins

ML approaches provide a promising way not only to predict phage virion proteins, but also their functions. Nami and collaborators [18] conducted a review on 10 different ML models for the previous purposes. The list of the models, alongside with the respective methods and features is presented in Table 1. Table 1 also shows diversity regarding ML methods and features, allowing flexibility when choosing the ML model(s) for the scope of this project. Nevertheless, models performances must be evaluated in the context of our work, to allow for a weighted decision of a model in detriment of another.

Table 1. ML models and respective methods and features. [18].

Predictor	Method
ANN	"ACC, protein isoelectric Points" + ANN
Naïve Bayes	"ACC, DPC" + CFS + Naïve Bayes
PVPred	g-gap DPC + ANOVA+SVM
PhagePred	g-gap DPC + ANOVA + Multinomial Naïve Bayes + ANN
PVP-SVM	"AAC, ATC, CTD, DPC, PCP" + RF-based feature selection + SVM
SVM-based	g-gap DPC + "ANOVA, mRMR" + SVM + ANN
Ensemble RF	"CTD, bi-profile Bayes, PseAAC, PSSM" + Relief + RF + ANN
Pred-BVP-Unb	CT, SAAC, bi-PSSM+SVM + ANN
PVPred-SCM	DPC + SCM + ANN
Meta-iPVP	Probabilistic feature+SVM + ANN

SCM scoring card method, *SVM* support vector machine, *AAC* amino acid composition, *ATC* atomic composition, *bi-PSSM* bi-profile position specific scoring matrix, *CTD* chain-transition-distribution, *CT* composition and translation, *DPC* dipeptide composition, *GDPC* g-gap dipeptide composition, *PCP* physicochemical properties, *SAAC* split amino acid composition.

In a recent work, Vieira and collaborators [27] developed the first ML tool to identify phage depolymerases. This tool is based on two algorithms, SVMs and ANNs that were used to train the machine learning models. This groundbreaking study shows immense progress in ML tools to identify phage lytic enzymes providing valuable insights for the future development of a ML tool to predict endolysins.

Bibliography

[1] Fatma Abdelrahman, Maheswaran Easwaran, Oluwasegun I. Daramola, Samar Ragab, Stephanie Lynch, Tolulope J. Oduselu, Fazal Mehmood Khan, Akomolafe Ayobami, Fazal Adnan, Eduard Torrents, Swapnil Sanmukh, and Ayman El-Shibiny. Phage-Encoded Endolysins. *Antibiotics*, 10(2):1–31, feb 2021.

[2] Yves Briers. Phage Lytic Enzymes. *Viruses*, 11(2), feb 2019.

[3] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60, jan 2015.

[4] Jesse Cahill and Ry Young. Phage Lysis: Multiple Genes for Multiple Barriers. *Advances in virus research*, 103:33, jan 2019.

[5] Luis F. Camarillo-Guerrero, Alexandre Almeida, Guillermo Rangel-Pineros, Robert D. Finn, and Trevor D. Lawley. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109.e9, feb 2021.

[6] Martha R.J. Clokie, Andrew D. Millard, Andrey V. Letarov, and Shaun Heaphy. Phages in nature. *Bacteriophage*, 1(1):31, jan 2011.

[7] Felipe H. Coutinho, Cynthia B. Silveira, Gustavo B. Gregoracci, Cristiane C. Thompson, Robert A. Edwards, Corina P.D. Brussaard, Bas E. Dutilh, and Fabiano L. Thompson. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, 8, jul 2017.

[8] Bjorn Criel, Steff Taelman, Wim Van Criekinge, Michiel Stock, and Yves Briers. PhaLP: A database for the study of phage lytic proteins and their evolution. *Viruses*, 13(7), jul 2021.

[9] Moïra B. Dion, Frank Oechslin, and Sylvain Moineau. Phage diversity, genomics and phylogeny. *Nature reviews. Microbiology*, 18(3):125–138, mar 2020.

[10] Iris Fernández-Ruiz, Felipe H. Coutinho, and Francisco Rodriguez-Valera. Thousands of Novel Endolysins Discovered in Uncultured Phage Genomes. *Frontiers in microbiology*, 9(MAY), may 2018.

[11] Robert D. Finn, Jody Clements, William Arndt, Benjamin L. Miller, Travis J. Wheeler, Fabian Schreiber, Alex Bateman, and Sean R. Eddy. HMMER web server: 2015 update. *Nucleic Acids Research*, 43(Web Server issue):W30, jul 2015.

[12] Hongquan Gou, Enhao Li, Yilun Xue, Yi Rong, Yihui Zhang, Cheng Chang, Wennan Guo, Shiyun Wang, Jingyang Tu, Chao Lv, Min Li, Jiewen Huang, Xiaokui Guo, Qingtian Li, and YongZhang Zhu. PhalydDB: An Extensive Phage-Derived Lytic Protein Database for Targeted Antimicrobial Engineering Design and Bacterial Host Prediction. *SSRN Electronic Journal*, jul 2022.

[13] Liliam K. Harada, Erica C. Silva, Welida F. Campos, Fernando S. Del Fiol, Marta Vila, Krystyna Dąbrowska, Victor N. Krylov, and Victor M. Balcão.

Biotechnological applications of bacteriophages: State of the art. *Microbiological Research*, 212-213:38–58, jul 2018.

[14] Graham F. Hatfull and Roger W. Hendrix. Bacteriophages and their genomes. *Current Opinion in Virology*, 1(4):298–303, oct 2011.

[15] Katarina Hojckova, Matej Stano, and Lubos Klucar. phiBIOTICS: catalogue of therapeutic enzybiotics, relevant research studies and practical applications. *BMC microbiology*, 13(1), 2013.

[16] Doug Hyatt, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, mar 2010.

[17] Elliot J. Lefkowitz, Donald M. Dempsey, Robert Curtis Hendrickson, Richard J. Orton, Stuart G. Siddell, and Donald B. Smith. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, 46(D1):D708–D717, jan 2018.

[18] Yousef Nami, Nazila Imeni, and Bahman Panahi. Application of machine learning in bacteriophage research. *BMC microbiology*, 21(1):193, jun 2021.

[19] Stephen Nayfach, David Páez-Espino, Lee Call, Soo Jen Low, Hila Sberro, Natalia N. Ivanova, Amy D. Proal, Michael A. Fischbach, Ami S. Bhatt, Philip Hugenholtz, and Nikos C. Kyrpides. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology 2021 6:7*, 6(7):960–970, jun 2021.

[20] Hugo Oliveira, Luís D. R. Melo, Sílvio B. Santos, Franklin L. Nóbrega, Eugénio C. Ferreira, Nuno Cerca, Joana Azeredo, and Leon D. Kluskens. Molecular aspects and comparative genomics of bacteriophage endolysins. *Journal of virology*, 87(8):4558–4570, apr 2013.

[21] David Paez-Espino, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. Uncovering Earth's virome. *Nature 2016 536:7617*, 536(7617):425–430, aug 2016.

[22] Mattia Pandolfo, Andrea Telatin, Gioele Lazzari, Evelien M. Adriaenssens, and Nicola Vitulo. MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems*, 7(5), oct 2022.

[23] Thomas Nordahl Petersen, Søren Brunak, Gunnar Von Heijne, and Henrik Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods 2011 8:10*, 8(10):785–786, sep 2011.

[24] Simon Roux, Steven J. Hallam, Tanja Woyke, and Matthew B. Sullivan. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, 4(JULY2015), jul 2015.

[25] Sonika Sharma, Soumya Chatterjee, Sibnarayan Datta, Rishika Prasad, Dharmendra Dubey, Rajesh Kumar Prasad, and Mohan G. Vairale. Bacteriophages and its applications: an overview. *Folia microbiologica*, 62(1):17–55, jan 2017.

[26] Dann Turner, Andrey N. Shkoporov, Cédric Lood, Andrew D. Millard, Bas E. Dutilh, Poliane Alfenas-Zerbini, Leonardo J. van Zyl, Ramy K.

Aziz, Hanna M. Oksanen, Minna M. Poranen, Andrew M. Kropinski, Jakub Barylski, J. Rodney Brister, Nina Chanisvili, Rob A. Edwards, François Enault, Annika Gillis, Petar Knezevic, Mart Krupovic, Ipek Kurtböke, Alla Kushkina, Rob Lavigne, Susan Lehman, Malgorzata Lobocka, Cristina Moraru, Andrea Moreno Switt, Vera Morozova, Jesca Nakavuma, Alejandro Reyes Muñoz, Jānis Rūmnieks, Bl Sarkar, Matthew B. Sullivan, Jumpei Uchiyama, Johannes Wittmann, Tong Yigang, and Evelien M. Adriaenssens. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, 168(2):1–9, feb 2023.

[27] Maria Vieira, José Duarte, Rita Domingues, Hugo Oliveira, and Oscar Dias. PhageDPO: Phage Depolymerase Finder. *bioRxiv*, page 2023.02.24.529883, feb 2023.

[28] Peter J. Walker, Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Evelien M. Adriaenssens, Poliane Alfenas-Zerbini, Donald M. Dempsey, Bas E. Dutilh, María Laura García, R. Curtis Hendrickson, Sandra Junglen, Mart Krupovic, Jens H. Kuhn, Amy J. Lambert, Małgorzata Łobocka, Hanna M. Oksanen, Richard J. Orton, David L. Robertson, Luisa Rubino, Sead Sabanadzovic, Peter Simmonds, Donald B. Smith, Nobuhiro Suzuki, Koenraad Van Doorslaer, Anne Mieke Vandamme, Arvind Varsani, and Francisco Murilo Zerbini. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, 167(11):2429–2440, nov 2022.

[29] Hongyu Wu, Jinjiang Huang, Hairong Lu, Guodong Li, and Qingshan Huang. GMEnzy: A Genetically Modified Enzybiotic Database. *PLoS ONE*, 9(8), aug 2014.

[30] Hongyu Wu, Hairong Lu, Jinjiang Huang, Guodong Li, and Qingshan Huang. EnzyBase: a novel database for enzybiotic studies. *BMC microbiology*, 12(1):54, dec 2012.