PHERI - Phage Host Exploration pipeline.

Andrej Baláž[1], Michal Kajsík[2,3], Jaroslav Budiš[1,2,4], Tomáš Szemeš[1,2,3], Ján Turňa[3]

[1]Geneton Ltd., Ilkovicova 8, 841 04 Bratislava, Slovakia

[2]Comenius University Science Park, Ilkovičova 8, 841 04 Bratislava, Slovakia

[3]Department od Molecular Biology, Comenius University Faculty of Natural Sciences, PRIF UK, Mlynská dolina, Ilkovičova 6, 842 15 Bratislava 4, Slovakia

[4]Slovak Centre of Scientific and Technical Information (SCSTI), Lamacska cesta 8/A, 811 04 Bratislava, Slovakia

Corresponding author: Michal Kajsík, e-mail: michal.kajsik@uniba.sk

Abstract

Antibiotic resistance is becoming a common problem in medicine, food, and industry, with multi-resistant bacterial strains occurring in all world regions. One of the possible future solutions is the use of bacteriophages in therapy. Bacteriophages are the most abundant form of life in the biosphere, so we can highly likely purify a specific phage against each target bacterium. The identification and consistent characterization of individual phages was a common form of phage work and included determining bacteriophages' host-specificity. With the advent of new modern sequencing methods, there was a problem with the detailed characterization of phages in the environment identified by metagenome analysis. The solution to this problem may be to use a bioinformatic approach in the form of prediction software capable of determining a bacterial host based on the phage whole-genome sequence. The result of our research is the machine learning algorithm based tool called PHERI. PHERI predicts suitable bacterial host genus for purification of individual viruses from different samples. Besides, it can identify and highlight protein sequences that are important for host selection. PHERI is available at https://hub.docker.com/repository/docker/andynet/pheri. The source code for the model training is available at https://github.com/andynet/pheri_preprocessing, and the source code for the tool is available at https://github.com/andynet/pheri.

Key words: *bacteriophages, machine learning algorithm, phage host determination.*

Introduction

Bacterial infections affect public health throughout human history. The introduction of antibiotics reduced human morbidity and mortality caused by infectious diseases dramatically. However, the emergence of multidrug-resistant pathogenic bacteria reverted the situation once again. Moreover, the problem of multi-drug resistance is getting worse. WHO calls attention to the infections, especially by *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, and *Neisseria gonorrhoeae*, and blood poisoning and foodborne diseases. These infections are becoming harder and sometimes nearly impossible to treat [1]. Moreover, antibiotic resistance is now recorded in every country [2]. One of the possible solutions is the use of bacteriophages in therapy. Phages have relatively simple structures composed of proteins (approx. 60%) that encapsulate a DNA or RNA genome (40%) [3,4]. Phages are among the most abundant entities in the biosphere, with an estimated 10^{31}-10^{32} phages in the world at any given time, moreover play a crucial role in regulating bacterial [1] populations; for example, phages are responsible for the death of approximately 20%-40% of all marine surface bacteria every 24 h [5–7]. They are ubiquitously and naturally distributed in all environments populated by bacterial hosts, including soil, water, air, and the intestines of humans and other animals [7–10]. The idea of using bacteriophages in therapy is not new. Phage therapy has been used in the former Soviet Union countries for decades [11,12], but in the last few years, it has begun to be applied in Western countries as well. Bacteriophages have proved their usefulness not only in animal models such as mice [13,14], cattle [15,16], chicken [17], zebrafish [18], or dog [19], but also when used in humans. Human phage therapy has gained reliance through research projects such as PhagoBurn [20] or practical experience in Georgia [21], leading to the first cases of phage use on patients in Western countries. In recent years, phage therapy has been successfully used for the intravenous treatment of bacterial infections in cystic fibrosis patients in the US and Georgia. It has been used against multi-drug resistant pathogens such as *Achromobacter xylosoxidans*, *Pseudomonas aeruginosa*, *Mycobacterium abscessus*, and *Burkholderia dolosa [22–25]*. Also, in the treatment of *Mycobacterium abscessus*, the patient was treated with a cocktail of three phages, of which only one was naturally lytic, and the other two were engineered to increase their lysis efficiency by deleting the receptor gene or its HTH domain [25]. These studies used well-characterized phages from collections, with known hosts, one of the primary conditions for their successful practical

use. Characterized phages were amplified on the host, purified, and subsequently sequenced. However, the introduction of high throughput sequencing allowed us to examine metagenomic colonies of bacteria right from the environment. This method has the potential to discover a considerable amount of new species, which were not cultivable before. However, it also produces more and more phage genomic sequence data without an identified host. Luckily, the host range of known phages tends to be relatively narrow, often consisting of only a subset of strains making up a single bacterial species [26]. The problem of unknown hosts can be alleviated by a bioinformatic approach. For a successful phage infection, it needs not only to adsorb to the host surface and insert its genetic information, but it also needs to overcome its immune response and ensure successful transcription and translation. It is therefore essential to consider changes in bacterial surface structures and thus in the presence of phage receptors on individual strains membranes within the species [27]. Equally important is the consideration of the bacterial host immune response as restriction-modification systems [28,29], CRISPR mechanism [30] or abortive systems [31,32]. Also important are factors affecting phage gene transcription and translation, such as the availability of specific tRNA or sufficient amino acids. However, the change in specificity may be due to overcoming the host response, as in obtaining the resistance of the CRISPR system [33]. All these parameters can negatively affect the host prediction. Nevertheless, several groups have already attempted to bioinformatically elucidate the phage-host interaction using a variety of approaches and tools such as Virsorter [34], MGTAXA [34,35] or HostPhinder [36]. Our goal was also to create a bioinformatic tool for predicting the host from the whole genome sequence, but we chose the machine learning algorithms approach. The use of machine learning algorithms has proved to be suitable for phage biology, as evidenced by their use in the search for phage virions [37], improved phage genome annotation [38] as well as phage classification [39,40]. Our pipeline, PHERI, re-annotates phage genomes, uses TRIBE-MCL for rapid and accurate clustering of annotated protein sequences [41,42] and binary decision tree classifier to predict phage host. The rationale behind our method lies in a close relationship between the genomic sequence of a gene and biological function of translated protein. Even if the function of the gene is unknown, the presence of similar sequences in the phages infecting the same hosts indicate that mentioned sequences are related to the host specificity. Presence of such sequences in the tested genome may resolve potential host.

Material and Methods

We used phage genome sequences collected from several publicly available databases to identify clusters of gene sequences that are specific for distinct hosts. We use them to classify unlabeled phage genomes using the following analysis steps. At first, coding regions of the genomic sequence are identified and extracted. Then, they are compared with gene sequences collected from phages with known hosts, resulting in the reduced binary vector representation of the phage. Finally, the vector is classified to identify potential hosts of the phage.

Collection of phage sequences

We downloaded genomic sequences of phages from three publicly available databases using automated in-house scripts; 6,091 records from GenBank [43], 2,070 records from ViralZone [44,45] and 2,567 records from PhageDB [46]. Although these databases cover the majority of currently sequenced and published phages, we made downloading step easily extensible for adding new sources of phage sequences that may emerge in the future. Downloaded records were highly redundant, mainly because many phage sequences were simultaneously presented in more databases. Therefore, we merged downloaded datasets, and removed duplicated records, resulting in a non-redundant dataset of 7,064 phage sequences capable to infect 183 bacterial genera (Figure 2). The hostname and taxonomy for each sequence were obtained and unified according to NCBI taxonomy to allow computer processing. The phages with hosts from the 50 most abundant bacteria species were selected for further analysis. The phages outside this group were discarded due to an insufficient number of samples for machine learning analysis. Genomes were further divided, using random sampling, into two distinct datasets; training set with 4,723 (80%) sequences and testing set with 1,202 (20%) sequences. The training set was utilized to identify clusters of common gene sequences and train parameters of the classifier. The accuracy of the method has been validated on the testing set (Table 1).

Extraction and annotation of genes

Phage genome sequences were annotated with locations of genes and their biological function. Although gene annotations of particular genomes are part of genomic records in the utilized

databases, we decided to annotate sequences from scratch. This way, we ensured consistency of annotations across our data with up-to-date knowledge. We used a publicly available pipeline called Prokka [47] to identify and annotate genes. First, coordinates of coding DNA sequences (CDS) were found with Prodigal tool [48]. After the locations of genes are predicted, Prokka can start to annotate functions of all CDSs. Prokka does this by comparison of a sequence to several databases of proteins with an experimentally determined function [45,49] or pre-processed protein families and domains [45,50,51]. The Prokka pipeline was run with the parameter **--kingdom Viruses** to take into account that we were annotating genes in phages.

Clustering of gene sequences

We compared extracted genes to identify clusters of highly similar sequences, presumably with the same biological function. Since thorough pairwise comparison of all proteins in the training set would be overly time-consuming, we employed a multi-step approach. In the first step, genes were deduplicated using CD-hit [52] to reduce the enormous number of sequences. Then, gene pairs with at least some local sequence similarity were identified using an optimized implementation of the Blast alignment tool [53], called CrocoBLAST [54]. Afterwards, the identified pairs with high local sequence similarity underwent thorough pairwise alignment [55] to retrieve more accurate similarity scores. Based on these similarity scores, we identified gene clusters using the Markov Cluster Algorithm implemented in package MCL [56]. We recovered 32,281 gene clusters. A substantial portion of clusters contained only a small number of gene sequences. We think such clusters do not have enough information to significantly help the classifier and only increase the chance of overfitting. Therefore, we removed all clusters with genes from less than 1% of phages from the training set. In the result, we obtained 1,965 clusters that were used further in the classification.

Training classification model

We trained a binary decision tree classifier [57] for each bacterial host from the dataset separately, since a united classifier for all potential hosts would be too complicated for coherent interpretation. Also, the phage sequence may be labelled with multiple bacterial hosts. The separate classification allows to label less-specific phages with multiple admissible hosts. At first, phage sequences from

the training set were transformed to the reduced integer vector representation, where value $a_{i,j}$ represents a number of genes from cluster j belonging to phage i. For each host, we trained a classifier to predict if an input vector represents a phage that can infect a given host. Each node in the resulting decision tree represents a single gene cluster with informative value regarding phage specificity. Presence or absence of such genes guides decision along the tree. Each informative cluster may be annotated with biological function to improve the interpretation of the decision process. Gene clusters without known function are good candidates for follow-up experimental evaluation.

Classifying novel sequence

A novel phage genome sequence is classified using similar steps. At first, gene sequences are identified using Prokka. Then, collected gene sequences are compared with the gene clusters using Blast. Genes with significant matches with any sequence from a cluster are assumed as members of clusters. Finally, the set of genes is transformed into a reduced vector representation. The vector is labelled with all trained classifiers. All bacteria with positive classification are assumed as potential hosts for the phage.

Bacterial strains and growth conditions

Escherichia, Cronobacter, Enterobacter, Salmonella, Klebsiella, Staphylococcus, Proteus, Morganella and *Citrobacter* strains were isolated from clinical or food samples in our laboratory or were obtained from the collections of Nottingham Trent University, UK, the Belgian Coordinated Collections of Microorganisms, the Czech Collection of Microorganisms or from Slovak Food Research Institute. We used a total of 33 strains of the genus *Escherichia, 26 Cronobacter, 15 Salmonella, 5 Enterobacter, 3 Proteus, 3 Staphylococcus, 2 Klebsiella, 2 Pseudomonas and one Morganella morganii* strain. We later expanded the set of strains by five bacteria from the genus *Citrobacter.* The Luria-Bertani (LB) broth and LB agar were the general-purpose media used to cultivate strains.

Isolation of bacteriophages

Bacteriophage DevCS701 was isolated from a sample from the Devínska Nová Ves wastewater treatment plant on the indicators *C. sakazakii* NTU701. Wastewater was sterilized by passage through a 22-µm filter and mixed with an equal volume of twofold concentrated LB medium and up to 1% of overnight bacterial culture. The inoculated mixture was cultivated overnight at 37 °C with shaking. Single-species phages were isolated by three repeated isolations from single plaques on double agar, followed by ultracentrifugation in a CsCl gradient [58].

Plaque assay and host range

The 200µl overnight bacterial culture was supplemented with 10 µl of 1 M CaCl2 and 10 µl of 1 M MgCl2, mixed with 5 ml of top agar (0.2% peptone, 0.7% NaCl and 0.7% agar), and poured onto an LB agar plate. 10 µl of the appropriate bacteriophage suspension (102-1010 PFU/ml) was spotted onto the plate and incubated overnight. Alternatively, 20 µl of bacteriophage suspension was mixed with 200 µl of overnight bacterial culture and with 5 ml of top agar and poured onto the LB agar. After overnight cultivation at 37 °C, the plaques were counted. The strain C.sakazakii NTU701 was used as a reference for determination of the efficiency of plating (EOP) for phage Dev-CS701.

Phage adsorption

One hundred eighty µl overnight bacterial culture (OD600 =1) was mixed with 20 µl of phage suspension (108 PFU/ml; the multiplicity of infection = 0.001) at 37 °C. After 10 min, 10µl of the sample was diluted in 0.99 ml of cold SM buffer (100 mM NaCl, 8 mM MgSO4,50 mM Tris-HCl, pH 7.5, 0.002% gelatin) and centrifuged. Unadsorbed phages from supernatants were counted by plaque assay, and the amount of phage adsorbed was calculated as the percentage of cell-bound phage. The measurements were repeated in triplicate.

Results

In this study, we developed and benchmarked PHERI, a tool for predicting the bacterial host species of phages. The method is based on the assumption that phages infecting the same host share similar protein sequences. In making predictions, PHERI relies on a reference database, in which sequences of phages have been annotated, and resulting genes have been categorized based

on similarity. Subsequently, host-specific clusters and cluster combinations were searched and created 50 family-specific decision trees. The genomes of the tested phages, for which the hosts should be predicted, are also annotated, categorized and compared to decision trees. If the test bacteriophage contains sequences similar to those in clusters in a specific tree, we assume that it can infect the host for which the decision tree is created. The prediction was then experimentally verified on phage samples from our collection.

Developing PHERI method

The method uses a reference database that we made of unique phage sequences from three databases - GenBank, ViralZone and PhagesDB. In the exploratory stage of our analysis, we used python library scikit-learn [59] for principal component analysis. Reduced representation of phages in the form of an integer matrix was used as an input. The first few principal components were used to create plots in the python library matplotlib. In Figure 1, we can see data visualized with the first principal component on the x-axis and the second principal component on the y-axis. Each data point represents one phage record, and the colour of the particular data point corresponds to the genus of that phage. Most phage records are located around the centre, with some distinct groups of Mycobacterium and Staphylococcus phages outside the centre. The first two principal components retained less than 21 % of the dataset variability. This result suggests that many predictors are contributing valuable information to the dataset. To take into account this high number of predictors and at the same time keep the interpretability of the resulting model acceptable, we choose Decision Tree Classifier as our model.

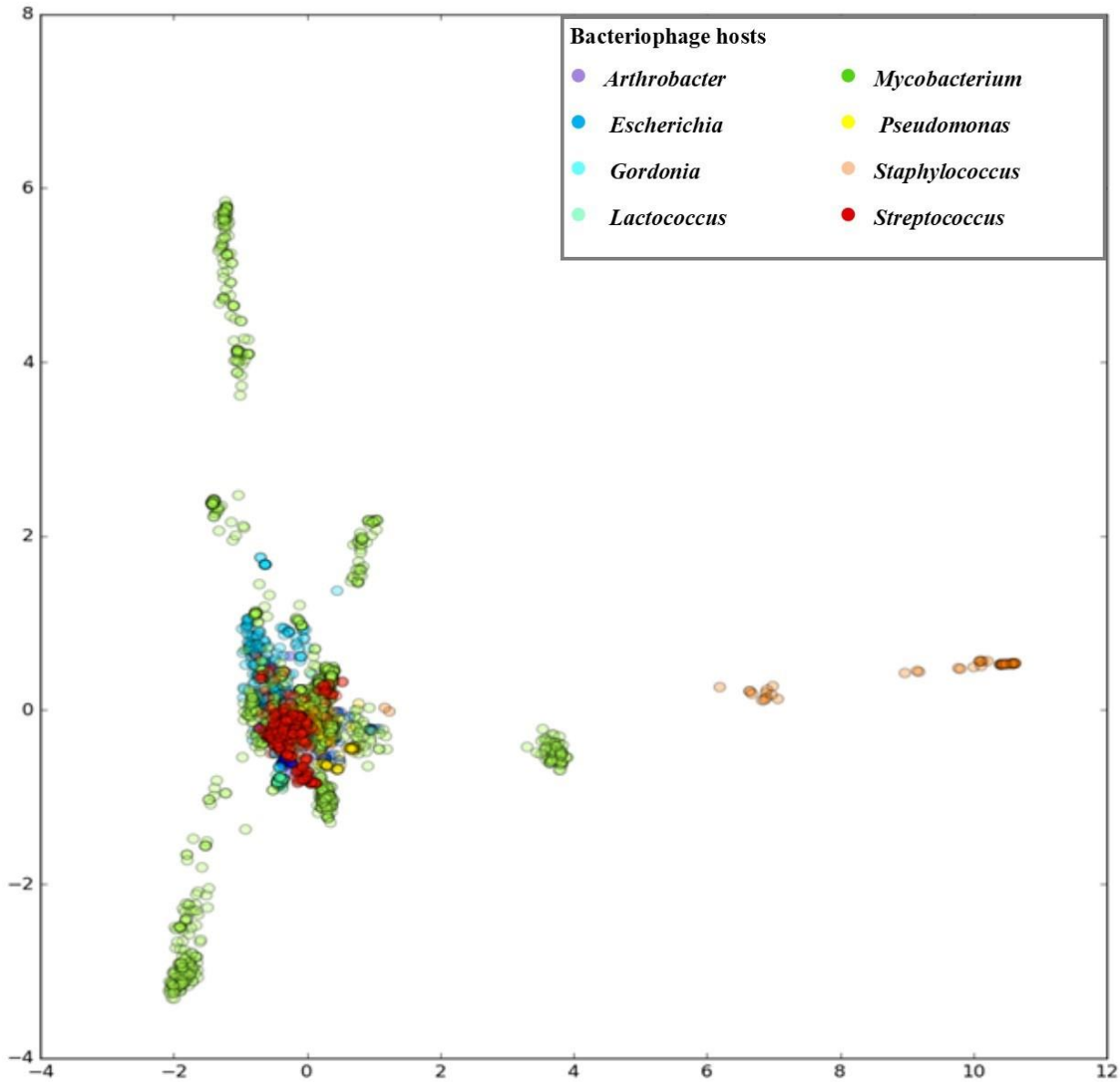

Figure 1: Principal component analysis. First two principal components, PC1 (11.57% of variability) on the x-axis, PC2 (9.19% of variability) on the y-axis

Before the feature selection, our training dataset consisted of a matrix with 4723 rows, representing phages and 32281 columns, representing gene clusters. The high dimensionality of our data could lead to the increased probability of overfitting of models on data. To address this concern, we decided to perform feature selection as a process of removing dimensions with low importance from the dataset. The reason to prefer feature selection over feature extraction methods as PCA presented in the previous section was that we wanted our tree models to be interpretable in terms

of important clusters rather than in terms of principal components. Because we expected many clusters with a small number of genes, our choice for feature selection method was the **Variance Threshold**. The Variance Threshold method is a simple method that removes columns with variance under a certain threshold. Using this technique, we removed columns in the matrix with ones in more than 99% of cases or with zeros in more than 99% of cases. The reduced matrix had 4723 rows and 1965 columns. In our work, we used a Decision Tree Classifier implementation from the python library scikit-learn. For each group of phages with hosts from selected genera, we created one model. Each of those models was trained to answer a question, whether one particular phage was able to infect bacteria from a specific genus. Models were trained with reduced matrix used as features. The ability to infect a specific genus was used as labels. To prevent overfitting of our trees, we also set the parameter **min_impurity_split** to *the value 0.03*. This parameter enabled a threshold for splitting leaves and therefore only nodes with an impurity index greater than 0.03 were divided. The threshold 0.03 was determined empirically. Lower values created a tree with many nodes, where the risk of overfitting was high and greater values did not have enough nodes to maintain a model's accuracy. With this approach, we created a model for each of our 50 selected host genera and visualized it with the python library Graphviz. For classification, we expected to have a complete sequence of bacteriophage.

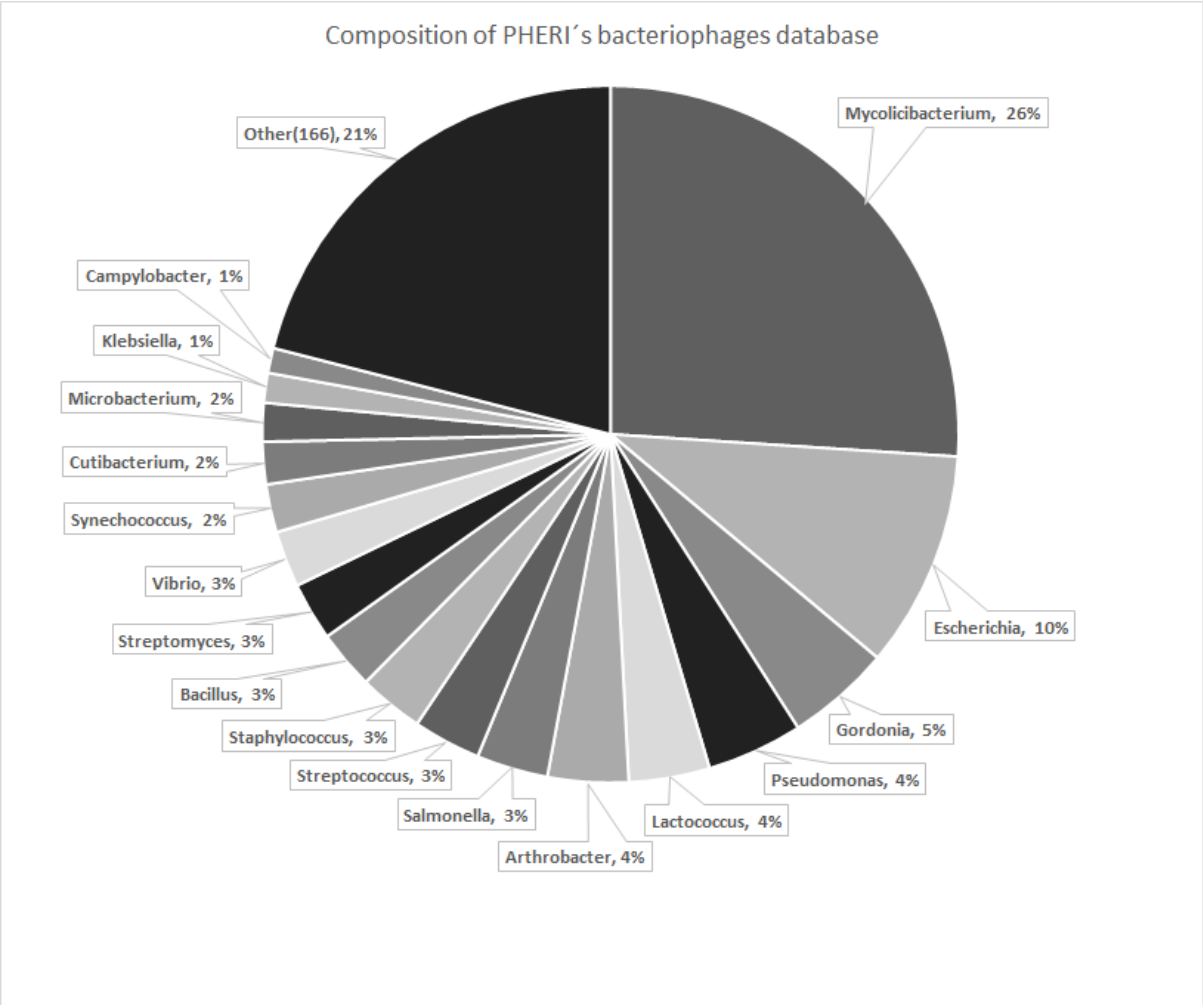

Figure 2: Composition of hosts infected by bacteriophages from the Pheri database. The database is made up of bacteriophages infecting at least one representative of 183 bacterial families.

Host prediction evaluation

To examine the accuracy of our models, we classified all bacteriophages from our test dataset. Test dataset contained 1,202 phage records. Resulting predictions were aggregated and the number of correctly predicted (TP + TN), false-positive (FP), and false-negative (FN), sensitivity (TPR = TP/P), specificity (TNR = TN/N) and informedness (BM = TPR + TNR - 1) was recorded. From the identified values, we determined the accuracy, sensitivity, specificity and informedness prediction for 50 bacterial genera with the highest number of infecting phages. PHERI predicted a host of bacteriophages infecting *Leuconostoc, Reugenia* and *Helicobacter* the best. Accuracy, sensitivity and specificity equal to or close to 100%. At the opposite end of the spectrum were bacteriophages infecting the genera *Stenotrophomonas* and *Citrobacter* (Table 1, Figure 3).

Test dataset n=1202	t_pos	t_neg	f_pos	f_neg		t_pos	t_neg	f_pos	f_neg
Leuconostoc	4	1198	0	0	Escherichia	82	1044	31	45
Ruegeria	3	1197	2	0	Enterococcus	6	1189	3	4
Helicobacter	6	1194	2	0	Listeria	4	1195	0	3
Paenibacillus	6	1192	4	0	Erwinia	5	1192	1	4
Cutibacterium	25	1173	4	0	Campylobacter	8	1180	7	7
Moraxella	7	1190	5	0	Salmonella	20	1147	13	22
Synechococcus	29	1166	7	0	Lactobacillus	5	1185	6	6
Lactococcus	47	1145	9	1	Clostridioides	2	1197	0	3
Streptococcus	39	1152	10	1	Clostridium	2	1195	2	3

Mycolicibacterium	320	847	33	2	Yersinia	2	1192	5	3
Staphylococcus	37	1155	8	2	Vibrio	13	1163	6	20
Arthrobacter	45	1150	4	3	Rhizobium	1	1198	1	2
Rhodococcus	11	1189	1	1	Klebsiella	5	1176	8	13
Microbacterium	21	1171	8	2	Xanthomonas	1	1194	3	4
Bacillus	32	1156	11	3	Cronobacter	1	1193	4	4
Gordonia	55	1135	5	7	Pectobacterium	1	1194	2	5
Flavobacterium	6	1194	1	1	Pseudoalteromonas	1	1192	4	5
Acinetobacter	9	1188	3	2	Brucella	1	1194	1	6
Pseudomonas	46	1129	16	11	Ralstonia	1	1193	2	6
Aeromonas	7	1190	3	2	Cellulophaga	1	1193	2	6
Corynebacterium	3	1194	4	1	Burkholderia	1	1191	4	6
Caulobacter	3	1193	5	1	Shigella	1	1184	7	10
Proteus	2	1199	0	1	Stenotrophomonas	0	1199	0	3
Mannheimia	2	1197	2	1	Citrobacter	0	1196	1	5
Streptomyces	23	1164	3	12	Mycobacterium	0	1196	4	2

Table 1: Number of true positive, true negative, false positive and false negative predictions on the testing dataset.

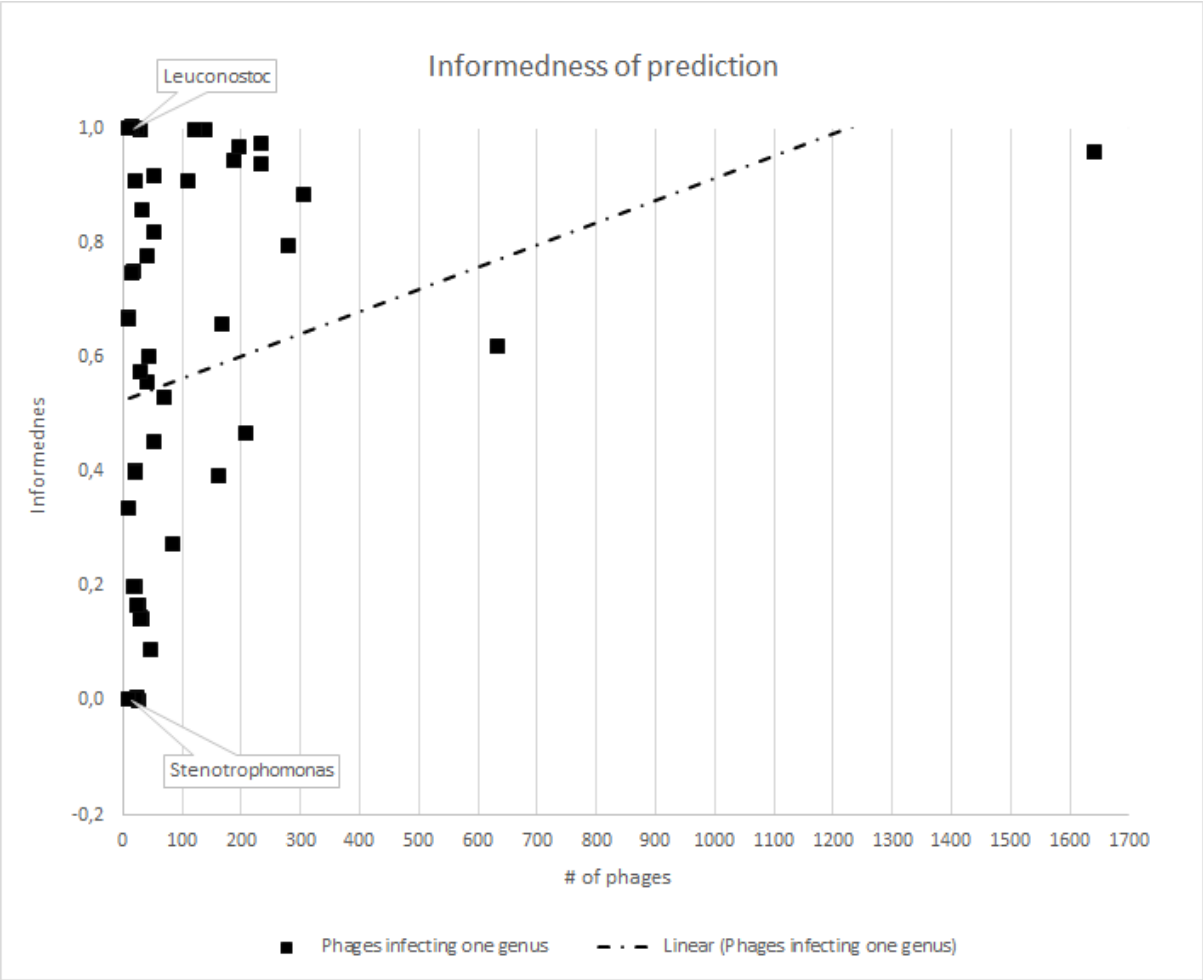

Figure 3: Informedness of PHERI host prediction. Each point represents phages infecting one bacterial genus. Value *Informedness* estimates the probability of an informed decision, the closer the values are to one, the more credible the prediction is. The trendline confirms the hypothesis that the informedness value increases with the growing number of specific phages.

Comparison of PHERI to other tools

Surprisingly, there are not many tools with the same goal as PHERI. We discuss some notable examples in the Discussion. Here we provide a comparison with the tool HostPhinder, where the goal is the most aligned with ours. We calculated the accuracy on the testing dataset, which was held out at the beginning of the analysis. This hold-out testing dataset mimics the conditions during

the use in practice in the laboratory, where the only information we have is the sequence of the isolated phage. The accuracy of both tools was remarkably high (>0.85). In Figure 4, we show the accuracy across all 50 selected genera, with the accuracy of HostPHinder on the x-axis, the accuracy of PHERI on the y-axis and the genera marked as blue points. PHERI achieved consistently high accuracy around 0.98 even for more challenging targets, where HostPhinder's accuracy dropped below 0.90.

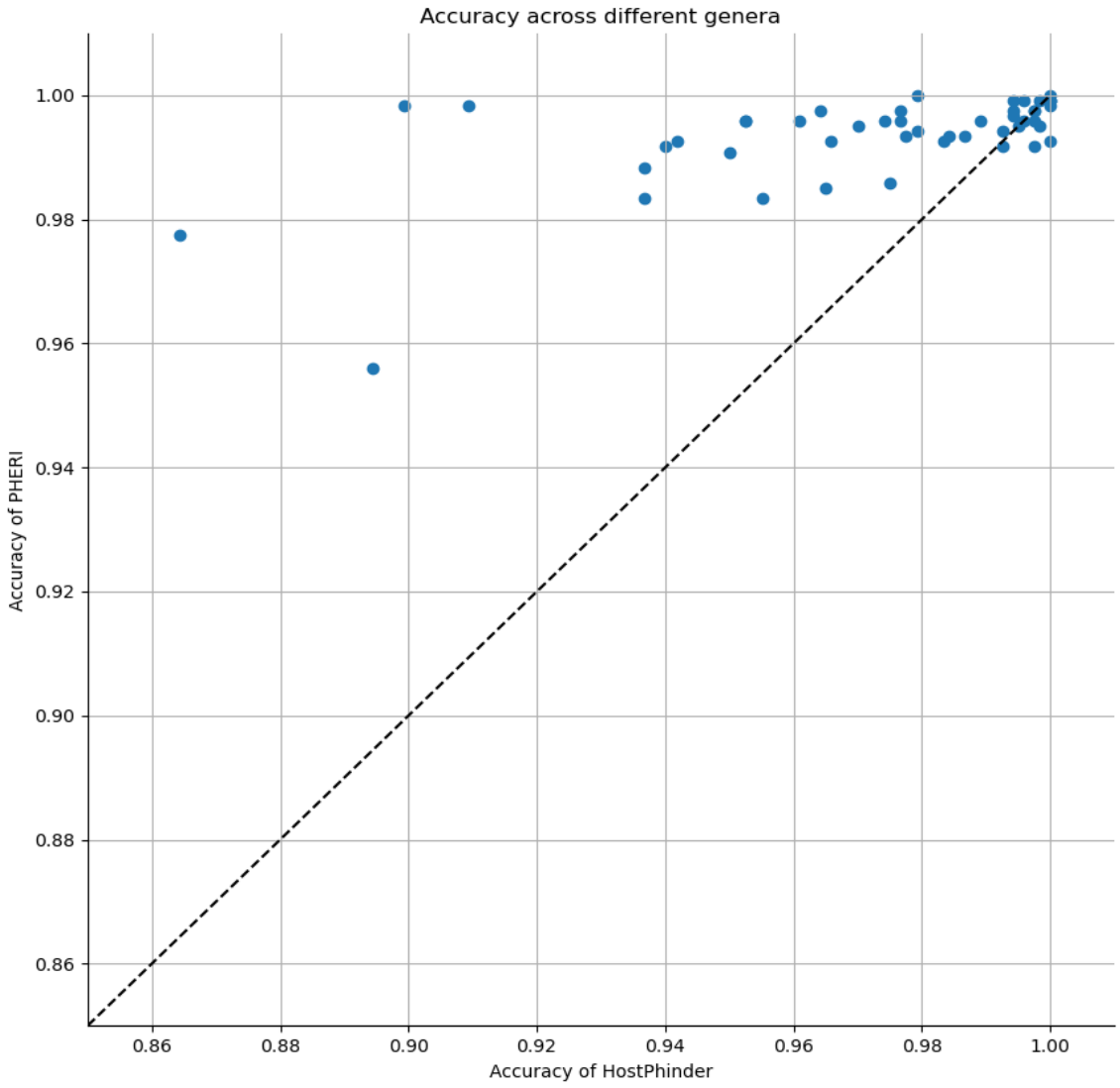

Figure 4: Accuracy of host prediction across 50 bacterial genera

Host prediction for new isolated bacteriophages

The PHERI's functionality was also verified by determining the host of phages isolated in our lab and not added to the public databases. Tested bacteriophages were isolated from wastewater from Bratislava, Slovakia. Their host specificity, as well as whole-genome sequence, was previously determined using standard wet science methods. The bacterial host genus for five out of six phages was successfully predicted using the PHERI method. Moreover, for the DevCS-701 phage, PHERI determined different bacterial genus (Table 2).

Bacteriophage	Closest relative	Real host	PHERI prediction
Dev-CS701	vB_CsaM_leB (KX431559.1)	*Cronobacter*	*Citrobacter*
vB_EcoM_VP1	vB_EcoM_JS09 (KF582788)	*E.coli*	*Escherichia*
vB-EcoM_KMB43	Rb49 like virus (AY343333)	*E.coli*	*Escherichia*
vB_KpnP_VP3	KPV811(KY000081)	*Klebsiella*	*Klebsiella*
vB_EcoP_VP5	64795_ec1(KU927499)	*E.coli*	*Escherichia*
PetSE1	vB_SenS-Ent1(NC_019539.1)	*Salmonella*	*Salmonella*

Table 2. Host prediction of newly isolated and sequenced phages.

According to laboratory tests, bacteriophage Dev-CS701 infects isolates from the genus *Cronobacter*, although PHERI predicted isolates from *Citrobacter* genus as the most likely candidate. For this reason, the host panel was expanded to include *Citrobacter* strains and re-established specificity. Extended host panel proved PHERI prediction since the Dev-CS701 phage infected a representative of the genus *Citrobacter*, namely *Citrobacter gillenii* CCM 4711.

However, the bacteriophage was not able to infect all *Citrobacter* strains. For this reason, we also examined the bacteriophage adsorption rate to the tested isolates (Figure 5). Dev-CS701 was able to bind to six out of seven *Cronobacter* strains and two out of four *Citrobacter* strains as well. Other strains did not reach high values, but the increased rate of adsorption on *Enterobacter* strains are also interesting.

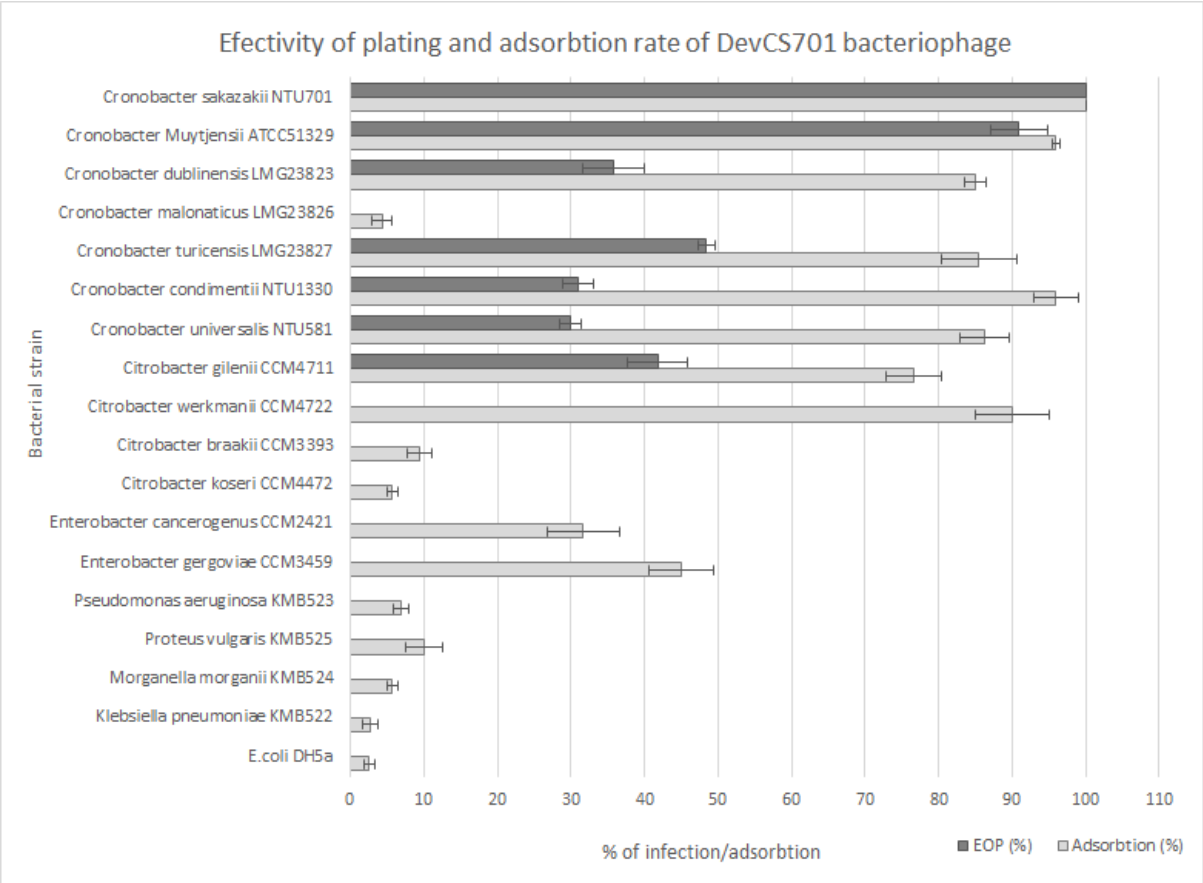

Figure 5: Effectivity of plating and adsorption rate of DevCS701 phage on various hosts.

Despite partial proof of the accuracy of the prediction, we decided to determine the cause of the selection of the genus *Citrobacter* instead of the genus *Cronobacter* by examining the decision trees. Phage contains sequences classified into clusters from both *Citrobacter* and *Cronobacter* decision trees, actually clusters 54 and 170 were found in both trees. In total, the phage had six

sequences that were classified in five clusters for both decision trees. However, PHERI was able to classify phage only according to the *Citrobacter* decision tree.

Discussion

The study of bacteriophages could help solve many of the problems with resistant bacteria in medicine, veterinary, food and other industries. One of the basic criteria for bacteriophage utilization is the knowledge of their whole genome sequence as well as the host [60,61]. The classic bacteriophage characterization methods were based on phage studies with a known host range and subsequent sequencing. However, since the advent of new massively parallel sequencing methods, these procedures are often reversed. It is also possible to identify bacteriophages that infect non-cultivable bacteria, the so-called "bacterial dark matter". Our studies have been previously focused on the identification of specific bacteriophages capable of infecting foodborne pathogens [62,63]. However, by using metagenomic data in our research, we discovered phages without the known host. Therefore, we developed a bioinformatic pipeline for predicting phage bacterial host genus from the whole-genome sequences, PHERI, based on machine learning algorithms. A couple of groups have already tested the idea of using a bioinformatic approach to identify phage hosts. For example, Martínez-Garcia et al. described one possibility of identifying a host without cultivation. They retrieved genomic content of individual cells from an environmental sample using single-cell genomic technologies, then hybridized against a set of phage genomes from the same sample, immobilized on a microarray and sequenced positive hybridization cases. Using this method, they were able to pinpoint viruses infecting the ubiquitous hyper-halophilic Nanohaloarchaeota, included in the so-called 'microbial dark matter' [64]. Another approach of the virus-host adaptation analysis was chosen by Roux et al. They developed a bioinformatics tool for virus sequence identification. VirSorter identified prophage sequences through a combination of detection of hallmarks viral genes, enrichment in viral-like genes, depletion in PFAM affiliated genes, enrichment in uncharacterized genes, enrichment in short genes and depletion in strand switching [34]. This tool was able to identify 12,498 virus-host linkages from almost 15 000 bacterial and archeal genomes. Identified prophage sequences came from 5492 microbial genomes, and provided first viral sequences for 13 new bacterial phyla. In their study, they also analyzed the virus-host adaptation in compositions in terms of nucleotide frequency and codon usage showing the strongest signal of adaptation to the host genome given by tetranucleotide frequency

(TNF)[65]. Another classification method to predict the taxonomy of bacterial hosts for uncharacterized viral metagenomic sequences, that does not rely on homology or sequence alignment, was developed by Williamson et al. In their study explaining the composition of the marine virome in the Indian Ocean, they also described the bioinformatic tool MGTAXA, which links phage sequences to the highest scoring bacterial taxonomic model based on polynucleotide genome composition similarity between the virus and host genomes [35]. An excellent tool for host prediction is also HostPhinder created by Villarroel et al. The HostPhinder is based on the assumption that genetically similar phages are likely to share bacterial hosts. The tool utilizes a phage database with known sequences that are divided into k-mers. Phages with an unknown host are also divided into k-mers and compared to a database. The high similarity of short DNA sequences between two phages will determine the likely host [36]. PHERI bases its prediction on machine learning technique, Decision Tree Classifier. The advantages of this technique are its interpretability and the potential for rapid improvement of prediction accuracy in the era of big datasets. The dependence of this technique on the size and quality of the dataset may also be considered its biggest drawback, as we showed for genera with small amounts of data. In this work, we used phage whole-genomics records from public databases with known hosts to create clusters of similar gene sequences that are specific to a certain genus. Sequences were annotated using Prokka [47], and genes were extracted using a custom script. Extracted genes were aligned using BLAST with a database of genes from the training set. Based on the retrieved similarity, clusters of genes were created. After that, a vector of integers was created for each phage representing genes and their corresponding assignment to clusters. This vector was passed to the decision tree model, and the resulting prediction was saved. For each of the 50 genera tested, a decision tree based on the necessity of specific clusters to infect the genus was created. Considering the mosaic structure of phage genomes, one of the advantages of using machine learning algorithms for phage host predictions is that only presence, absence and quantity of genetic elements influence the outcome. Thus, differences in locating and organizing individual genes do not affect the outcome of the pipeline prediction. In the evaluation test consisting of 1202 phages from the database, PHERI performed well when it reached the accuracy of 99.37% for the host genus prediction. However, the differences between bacterial genera were considerable as some hosts were easier to predict than others. We noticed a more accurate prediction of host genera with more than 200 phages in the database. The average sensitivity of prediction here was 81,8%. The prediction was

less sensitive for families with more than 100 and more than 10 phages, reaching 74,4 and 50.7%, respectively. The data shows that more representatives in the database increase the accuracy of the prediction. This result is probably due to the greater number of different host-specific protein sequences that PHERI clustered and incorporated into the decision tree. A greater number of clusters in the tree reduces the likelihood of incorrect prediction in the case of a phage with different mechanisms of infection. Chibani et al. has already described improvements in the prediction of machine learning algorithms based on the number of phages in the database in their phage classification study [39]. The small number of phages that infect individual species was the main reason why we designed PHERI to identify genera. In this way, we were able to increase the accuracy of the prediction and thus allow narrowing the range of hosts for later wet science host specificity tests. At the same time, we assume that by increasing the number of specific phages in the database, PHERI has the potential not only to increase the accuracy of genus prediction but also to predict the host at the species level. The number of specific phages in the database was not the only factor affecting the accuracy of the prediction. In particular, PHERI has identified all phages infecting the genus *Leuconostoc*, which has only 17 specific phages in the database. In contrast, in the case of the genus *Stenotrophomonas* with 13 phages in the database, none could be identified. By comparing bacteriophages infecting *Leuconoctoc*, we found that Leuconostoc-specific phages form two highly related groups belonging to the genera *Limdunavirus* and *Unaquatrovirus* within the family *Siphoviridae* and subfamily *Mccleskeyvirinae*. Homologous phages have probably a similar mechanism of infection that provides similar proteins. A similar conclusion was reached by Kot et al. in comparative genomic analysis of Leuconostoc phages [66]. PHERI, therefore, constructs a decision tree for a group of genetically-related phages easier and does not need a large number of viruses in the database. Similar results were obtained with the prediction of phage hosts of other bacterial genera with a small number of specific viruses. For example, the genera *Paenibacillus* or *Ruegenia* with 26 and 12 genetically related viruses, achieved a sensitivity of prediction over 99%. By contrast, phages infecting *Stenotrophomonas* are not genetically related, since some such as IME13(NC_029000.1) phage belong to the *Myoviridae* family[67], phage vB_SmaS_DLP_5(NC_042082.1) to the *Siphoviridae* family, or phages such as PSH1 (NC_010429.1) to the *Inoviridae* family[68]. Narrow and variable group of phages in the training set does not allow to construct a reliable decision tree.

Another factor that could affect the accuracy of the prediction is the ability of bacteriophages to infect bacteria of different species, even genera. Especially in cases of genetically related bacterial genera, several instances of phages with the ability to infect multiple genera have been described [63,69,70]. Even in these cases of known cross genera host specificity, only one genus name is found in the database.

We have used PHERI to locate a host of several phages isolated and characterized in our laboratory. The phages had established host specificity for strains of the genera *Escherichia, Cronobacter, Enterobacter, Salmonella, Klebsiella, Staphylococcus, Proteus, and Morganella.* Pheri correctly identified the host genus for five out of six phages (Table 2). In the case of the phage Dev-CS701, which infected strains of the genus *Cronobacter*, it predicted as a suitable host the bacteria from genus *Citrobacter*. Subsequent extended host specificity tests against strains of the genus *Citrobacter* confirmed that the phage also infected *Citrobacter gillenii* CCM 4711. Unfortunately, the phage was unable to form plaques on other strains of the genus. We have therefore tested the bacteriophage ability to recognize the bacterial surface, which confirmed that Dev_CS701 recognizes the surface not only of *C. gillenii* but also of *C. werkmanii*. Besides, a comparison of whole-genome sequences of phages by BLAST showed that Dev_CS701, in addition to the closest relative cronophage vB_CsaM_IeB (KX431559.1) [71], achieved similarity of over 96% to the Citrobacter-specific phages Margaery (KT381880.1) and Maroon (MH823906.1). Unfortunately, the detailed host specificity of the closest related phages is not yet publicly available. With our tool, we wanted to show a new way in the prediction of phage hosts. PHERI can help isolate live viruses from samples in wet labs by narrowing the range of possible hosts. There is also the potential to refine the prediction with the increasing number of new phages in databases. One of the advantages of host prediction based on the clustering of individual genes is the possibility of highlighting genes with unknown function necessary for infection. The identification of such genes may, in the future, help scientists to elucidate the mechanisms of infection of individual bacteriophages.

Conclusion

The importance of bacteriophages as a research subject is rising mainly due to the decreasing effectiveness of antibiotic treatments. Bacteriophages could be used as novel weapons

in the fight against bacterial infection. The goal of this work was to examine relationships between bacteriophage genomes and their bacterial hosts. The tool created is proof of the concept that the machine learning algorithm based tool can be used to search for bacterial hosts for viruses. PHERI can accelerate the isolation of individual bacteriophages from samples by narrowing the range of potential hosts. It also provides the user with information about likely proteins that are required for a successful infection.

Acknowledgement

This work was supported by Slovak Research and Development Agency under the contract No APVV-17-0526 and by the European H2020 Programme, Topic: MSCA-RISE-2019 872539 - PANGAIA – Pan-genome Graph Algorithms and Data Integration

References

1. Ogawara H. Comparison of Antibiotic Resistance Mechanisms in Antibiotic-Producing and Pathogenic Bacteria. Molecules. 2019;24. doi:10.3390/molecules24193430

2. Wilson ME. Antibiotics: What Everyone Needs to Know®. What Everyone Needs to Know(r); 2019.

3. Williamson KE, Fuhrmann JJ, Wommack KE, Radosevich M. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. Annu Rev Virol. 2017;4: 201–219.

4. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536: 425–430.

5. Wittebole X, De Roock S, Opal SM. A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens. Virulence. 2014;5: 226–235.

6. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. 2007;5: 801–812.

7. Yu P, Mathieu J, Lu GW, Gabiatti N, Alvarez PJ. Control of Antibiotic-Resistant Bacteria in Activated Sludge Using Polyvalent Phages in Conjunction with a Production Host. Environmental Science & Technology Letters. 2017. pp. 137–142. doi:10.1021/acs.estlett.7b00045

8. Simmonds P, Aiewsakun P. Virus classification – where do you draw the line? Archives of Virology. 2018. pp. 2037–2046. doi:10.1007/s00705-018-3938-z

9. Yu P, Mathieu J, Li M, Dai Z, Alvarez PJJ. Isolation of Polyvalent Bacteriophages by Sequential Multiple-Host Approaches. Appl Environ Microbiol. 2016;82: 808–815.

10. Ye M, Sun M, Huang D, Zhang Z, Zhang H, Zhang S, et al. A review of bacteriophage therapy for pathogenic bacteria inactivation in the soil environment. Environ Int. 2019;129: 488–496.

11. Chanishvili N. Phage therapy--history from Twort and d'Herelle through Soviet experience to current approaches. Adv Virus Res. 2012;83: 3–40.

12. Myelnikov D. An Alternative Cure: The Adoption and Survival of Bacteriophage Therapy in the USSR, 1922-1955. J Hist Med Allied Sci. 2018;73: 385–411.

13. Anand T, Virmani N, Kumar S, Kumar Mohanty A, Pavulraj S, Ch Bera B, et al. Phage Therapy for treatment of virulent Klebsiella pneumoniae infection in mouse model. J Glob Antimicrob Resist. 2019. doi:10.1016/j.jgar.2019.09.018

14. Dissanayake U, Ukhanova M, Moye ZD, Sulakvelidze A, Mai V. Bacteriophages Reduce Pathogenic Counts in Mice Without Distorting Gut Microbiota. Front Microbiol. 2019;10: 1984.

15. Smith HW, Huggins MB. Effectiveness of phages in treating experimental Escherichia coli diarrhoea in calves, piglets and lambs. J Gen Microbiol. 1983;129: 2659–2675.

16. Smith HW, Huggins MB, Shaw KM. Factors influencing the survival and multiplication of bacteriophages in calves and in their environment. J Gen Microbiol. 1987;133: 1127–1135.

17. Carrillo CL, Loc Carrillo C, Atterbury RJ, El-Shibiny A, Connerton PL, Dillon E, et al. Bacteriophage Therapy To Reduce Campylobacter jejuni Colonization of Broiler Chickens. Applied and Environmental Microbiology. 2005. pp. 6554–6563. doi:10.1128/aem.71.11.6554-6563.2005

18. Cafora M, Deflorian G, Forti F, Ferrari L, Binelli G, Briani F, et al. Phage therapy against Pseudomonas aeruginosa infections in a cystic fibrosis zebrafish model. Scientific Reports. 2019. doi:10.1038/s41598-018-37636-x

19. Marza JAS, Soothill JS, Boydell P, Collyns TA. Multiplication of therapeutically administered bacteriophages in Pseudomonas aeruginosa infected patients. Burns. 2006;32: 644–646.

20. Jault P, Leclerc T, Jennes S, Pirnay JP, Que Y-A, Resch G, et al. Efficacy and tolerability of a cocktail of bacteriophages to treat burn wounds infected by Pseudomonas aeruginosa (PhagoBurn): a randomised, controlled, double-blind phase 1/2 trial. Lancet Infect Dis. 2019;19: 35–45.

21. Zhvania P, Hoyle NS, Nadareishvili L, Nizharadze D, Kutateladze M. Phage Therapy in a 16-Year-Old Boy with Netherton Syndrome. Frontiers in Medicine. 2017. doi:10.3389/fmed.2017.00094

22. Hoyle N, Zhvaniya P, Balarjishvili N, Bolkvadze D, Nadareishvili L, Nizharadze D, et al. Phage therapy against Achromobacter xylosoxidans lung infection in a patient with cystic fibrosis: a case report. Research in Microbiology. 2018. pp. 540–542. doi:10.1016/j.resmic.2018.05.001

23. Law N, Logan C, Yung G, Furr C-LL, Lehman SM, Morales S, et al. Successful adjunctive use of bacteriophage therapy for treatment of multidrug-resistant Pseudomonas aeruginosa infection in a cystic fibrosis patient. Infection. 2019. pp. 665–668. doi:10.1007/s15010-019-01319-0

24. Aslam S, Courtwright AM, Koval C, Lehman SM, Morales S, Furr C-LL, et al. Early clinical experience of bacteriophage therapy in 3 lung transplant recipients. Am J Transplant. 2019;19:

2631–2639.

25. Dedrick RM, Guerrero-Bustamante CA, Garlena RA, Russell DA, Ford K, Harris K, et al. Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nat Med. 2019;25: 730–733.

26. Hyman P, Abedon ST. Bacteriophage Host Range and Bacterial Resistance. Advances in Applied Microbiology. 2010. pp. 217–248. doi:10.1016/s0065-2164(10)70007-1

27. Tu J, Park T, Morado DR, Hughes KT, Molineux IJ, Liu J. Dual host specificity of phage SP6 is facilitated by tailspike rotation. Virology. 2017;507: 206–215.

28. Hutinet G, Kot W, Cui L, Hillebrand R, Balamkundu S, Gnanakalai S, et al. 7-Deazaguanine modifications protect phage DNA from host restriction systems. Nat Commun. 2019;10: 5442.

29. Furi L, Crawford LA, Rangel-Pineros G, Manso AS, De Ste Croix M, Haigh RD, et al. Methylation Warfare: Interaction of Pneumococcal Bacteriophages with Their Host. J Bacteriol. 2019;201. doi:10.1128/JB.00370-19

30. Modell JW, Jiang W, Marraffini LA. CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature. 2017;544: 101–104.

31. Chopin M-C, Chopin A, Bidnenko E. Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 2005;8: 473–479.

32. Chen B, Akusobi C, Fang X, Salmond GPC. Environmental T4-Family Bacteriophages Evolve to Escape Abortive Infection via Multiple Routes in a Bacterial Host Employing “Altruistic Suicide” through Type III Toxin-Antitoxin Systems. Frontiers in Microbiology. 2017. doi:10.3389/fmicb.2017.01006

33. Stanley SY, Maxwell KL. Phage-Encoded Anti-CRISPR Defenses. Annu Rev Genet. 2018;52: 445–464.

34. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015. p. e985. doi:10.7717/peerj.985

35. Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, et al. Metagenomic exploration of viruses throughout the Indian Ocean. PLoS One. 2012;7: e42047.

36. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, et al. HostPhinder: A Phage Host Prediction Tool. Viruses. 2016;8. doi:10.3390/v8050116

37. Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front Microbiol. 2018;9: 476.

38. Salisbury A, Tsourkas PK. A Method for Improving the Accuracy and Efficiency of Bacteriophage Genome Annotation. Int J Mol Sci. 2019;20. doi:10.3390/ijms20143391

39. Chibani CM, Meinecke F, Farr A, Dietrich S, Liesegang H. ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks. doi:10.1101/558171

40. Lopes A, Tavares P, Petit M-A, Guérois R, Zinn-Justin S. Automated classification of tailed bacteriophages according to their neck organization. BMC Genomics. 2014. p. 1027. doi:10.1186/1471-2164-15-1027

41. Enright AJ. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002. pp. 1575–1584. doi:10.1093/nar/30.7.1575

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990. pp. 403–410. doi:10.1016/s0022-2836(05)80360-2

43. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic Acids Res. 2018;46: D41–D47.

44. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423.

45. Griffith M, Griffith OL. RefSeq (the Reference Sequence Database). Dictionary of Bioinformatics and Computational Biology. 2004.

46. Cosma CL, Sherman DR, Ramakrishnan L. The secret lives of the pathogenic mycobacteria.

Annu Rev Microbiol. 2003;57: 641–676.

47. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

48. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119.

49. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43: D204–12.

50. Bateman A. The Pfam Protein Families Database. Nucleic Acids Res. 2000;28: 263–266.

51. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, et al. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. 2001;29: 41–43.

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

53. Altschul S. Basic Local Alignment Search Tool. J Mol Biol. 1990;215: 403–410.

54. Tristão Ramos RJ, de Azevedo Martins AC, da Silva Delgado G, Ionescu C-M, Ürményi TP, Silva R, et al. CrocoBLAST: Running BLAST efficiently in the age of next-generation sequencing. Bioinformatics. 2017;33: 3648–3651.

55. Heringa J. Needleman-Wunsch Algorithm. Dictionary of Bioinformatics and Computational Biology. 2004.

56. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30: 1575–1584.

57. Oliphant TE. Python for Scientific Computing. Comput Sci Eng. 2007;9: 10–20.

58. Boulanger P. Purification of bacteriophages and SDS-PAGE analysis of phage structural

proteins from ghost particles. Methods Mol Biol. 2009;502: 227–238.

59. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

60. Mahony J, McAuliffe O, Paul Ross R, van Sinderen D. Bacteriophages as biocontrol agents of food pathogens. Current Opinion in Biotechnology. 2011. pp. 157–163. doi:10.1016/j.copbio.2010.10.008

61. Expert round table on acceptance and re-implementation of bacteriophage therapy. Silk route to the acceptance and re-implementation of bacteriophage therapy. Biotechnol J. 2016;11: 595–600.

62. Kajsík M, Bugala J, Kadličeková V, Szemes T, Turňa J, Drahovská H. Characterization of Dev-CD-23823 and Dev-CT57, new Autographivirinae bacteriophages infecting Cronobacter spp. Arch Virol. 2019;164: 1383–1391.

63. Kajsík M, Oslanecová L, Szemes T, Hýblová M, Bilková A, Drahovská H, et al. Characterization and genome sequence of Dev2, a new T7-like bacteriophage infecting Cronobacter turicensis. Arch Virol. 2014;159: 3013–3019.

64. Martínez-García M, Santos F, Moreno-Paz M, Parro V, Antón J. Unveiling viral-host interactions within the “microbial dark matter.” Nat Commun. 2014;5: 4542.

65. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. 2015;4. doi:10.7554/eLife.08490

66. Kot W, Hansen LH, Neve H, Hammer K, Jacobsen S, Pedersen PD, et al. Sequence and comparative analysis of Leuconostoc dairy bacteriophages. Int J Food Microbiol. 2014;176: 29–37.

67. Fan H, Huang Y, Mi Z, Yin X, Wang L, Fan H, et al. Complete Genome Sequence of IME13, a Stenotrophomonas maltophilia bacteriophage with large burst size and unique plaque polymorphism. J Virol. 2012;86: 11392–11393.

68. Liu J, Liu Q, Shen P, Huang Y-P. Isolation and characterization of a novel filamentous phage

from Stenotrophomonas maltophilia. Archives of Virology. 2012. pp. 1643–1650. doi:10.1007/s00705-012-1305-z

69. McCutcheon J, Peters D, Dennis J. Identification and Characterization of Type IV Pili as the Cellular Receptor of Broad Host Range Stenotrophomonas maltophilia Bacteriophages DLP1 and DLP2. Viruses. 2018. p. 338. doi:10.3390/v10060338

70. Peters DL, Lynch KH, Stothard P, Dennis JJ. The isolation and characterization of two Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. BMC Genomics. 2015. doi:10.1186/s12864-015-1848-y

71. Endersen L, Buttimer C, Nevin E, Coffey A, Neve H, Oliveira H, et al. Investigating the biocontrol and anti-biofilm potential of a three phage cocktail against Cronobacter sakazakii in different brands of infant formula. Int J Food Microbiol. 2017;253: 1–11.