PHERI - Phage Host Exploration tool.

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 health care, veterinary medicine, agriculture or food industry. Multi-resistant bacterial strains occur in all regions of the world. 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 it is highly likely that we can purify a specific phage against each target bacterium. A standard identification and consistent characterization of individual bacteriophages include host-specificity of viruses. Unfortunately, these routine methods are also considerably time consuming. With the advent of new modern sequencing methods, scientists are able to obtain multiple phage sequences from samples and identify more phages. However there appeared a problem with unknown host specificity of identified phages. The solution to this problem may be to use a bioinformatic approach in the form of prediction software capable to determine 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. In addition, the tool can identify and highlight protein sequences that are important for host selection.

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 situation of multidrug resistance is getting worse. WHO calls attention to the infections especially by *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, and *Neisseria gonorrhoeae*, blood poisoning and foodborne diseases, where 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]. Bacteriophages are 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 populations, for example, phages are responsible for the death of approximately 20%-40% of all marine surface bacteria every 24 h [5, 6] [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 countries of the former Soviet Union 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 human. 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, the phage therapy has been successfully used for the intravenous treatment of bacterial infections in cystic fibrosis patients in the US and Georgia and has been used against multi-drug resistant pathogens such as *Achromobacter*

xylosoxidans, Pseudomonas aeruginosa, Mycobacterium abscessus and *Burkholderia dolosa [22] [23] [24] [25]*. In addition, in the treatment of *Mycobacterium abscessus*, the patient was treated with a cocktail of three phages, of which one was naturally lytic, but the other two were engineered to increase their lysis efficiency by deleting the receptor gene or its HTH domain [25]. All these studies used well-characterized phages from collections, with known host, which is one of the basic conditions for their successful practical use. Most of the recently 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 or viruses right from the environment. This method has the potential to discover a huge 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 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 solved or at least alleviated by a bioinformatic approach. Successful use of a bioinformatic approach can be challenging. For a successful phage infection, it need not only 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 important to remember changes in a bacterial surface structures and thus in the presence of phage receptors on individual strains membranes within the species [27]. Equally important is the perception 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 a customary specificity may also be due to overcoming the host response, such 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 genus. 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

Collection of phage sequences

We downloaded genomic sequences of phages from three publicly available databases using automated in-house scripts. Database consisted of 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 a lot of phage sequences were simultaneously presented in more databases. Therefore, we merged downloaded datasets together, and removed duplicated records resulting in a non-redundant dataset of 7,064 phage sequences capable to infect 183 bacterial genera (Fig.1). 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 into two distinct datasets; the training set with 4,723 (80%) sequences and the testing set with 1,202 (20%) sequences. Sequences in the training set were utilized to identify clusters of common gene sequences and train parameters of the classifier. The accuracy of the method has been validated on sequences from the testing set (Tab. 1).

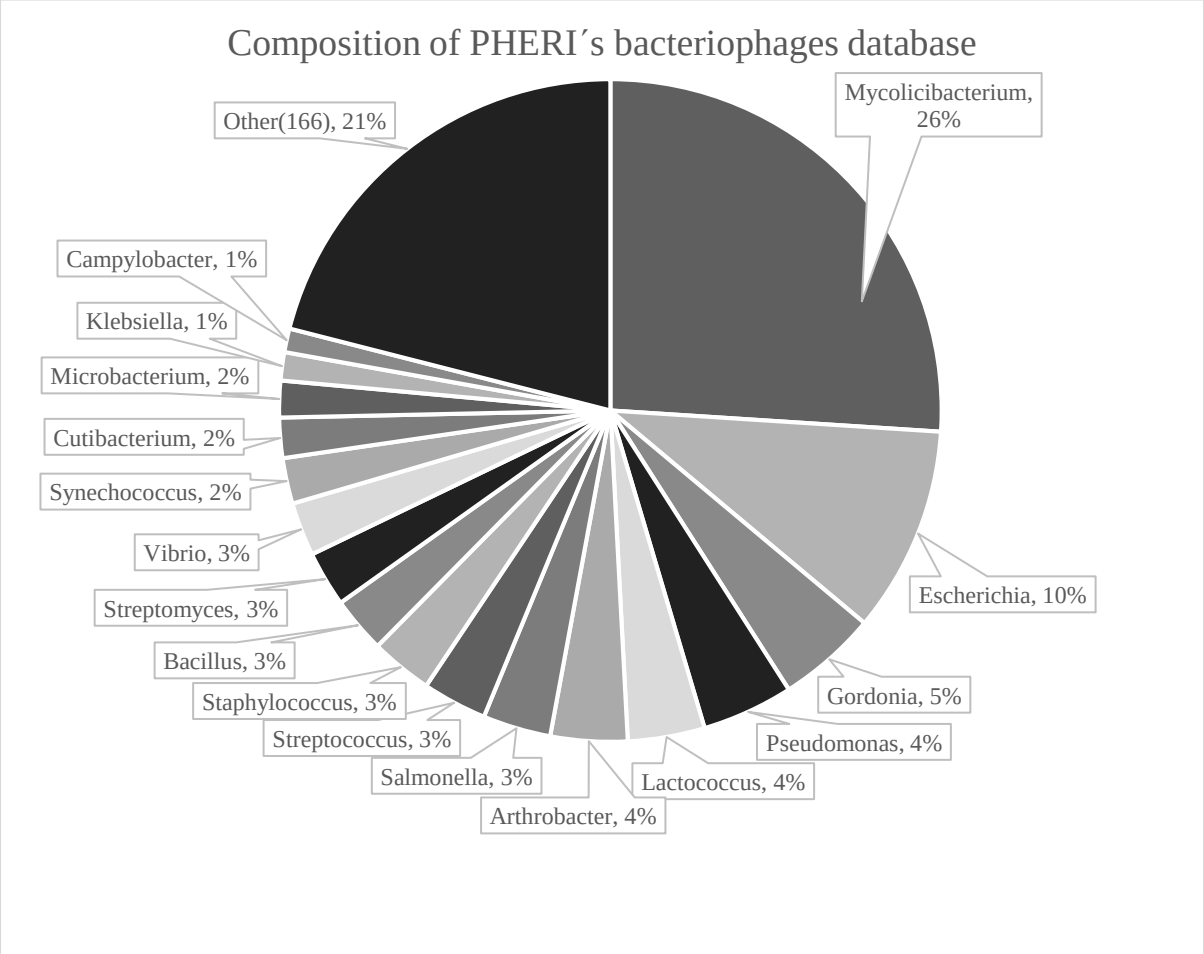

Figure 1:

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 genera

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 used 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 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. This is usually done through comparison of a sequence to several databases of sequences with an experimentally determined function [45, 49] or pre-processed protein families and domains [45, 50, 51].

Clustering of gene sequences

We compared extracted genes to identify clusters of recurrent sequences, presumably with the same biological function. Since thorough pairwise comparison of all sequences in the training set would be overly time consuming, we employed a two-step heuristic approach. At first, genome pairs with at least some local sequence similarity were recovered using an optimized implementation of the Blast alignment tool [52], called CrocoBLAST [53]. Only they underwent thorough pairwise alignment [54] to retrieve similarity scores. The rest of the pairs without any significant local similarity were scored with the lowest assumed similarity value. Based on summarized similarity scores we identified sequence clusters using the Markov Cluster Algorithm implemented in package MCL [55]. We recovered 32,281 gene clusters. A substantial portion of clusters was represented by a small number of gene sequences. We, therefore, removed all clusters present in less than 1% of phages as these clusters do not contain enough information to greatly help the classifier. 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 [56] for each bacterial host from the dataset separately, since united classifier for all potential hosts was too complex for coherent interpretation. In addition, 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 given host. Each node in the resulting decision tree represents a single gene cluster with informative value regarding a phage specificity. Presence or absence of such gene 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 phage 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, genes are 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

All bacterial strains used in this study 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 the Slovak Food Research Institute. Luria-Bertani (LB) broth and LB agar were the general-purpose media used to cultivate strains.

Isolation of bacteriophages

Bacteriophages vB_EcoM_VP1, vB-EcoM_KMB43, vB_KpnP_VP3, vB_EcoP_VP5, PetSE1 and Dev-CS701, were isolated from a wastewater samples from wastewater treatment plants in Bratislava, Slovakia. 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 the CsCl gradient.[57]

The 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 (10^2-10^{10} 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.

The Phage adsorption

One 180 µl overnight bacterial culture (OD600 =1) was mixed with 20 µl of phage suspension (10^8 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 $MgSO_4$, 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

Developing PHERI method

The method uses a reference database that we made from unique phage sequences of publicly available databases GenBank, ViralZone and PhagesDB. The python library scikit-learn [58] was used for principal component analysis. Reduced representation of phages in the form of a binary matrix was used as an input. First few principal components were used to create plots in python library matplotlib. In the Figure 2 we can see data visualized with the principal component one on the x-axis and the principal component two on the y-axis. Each data point represents one phage record and the color of the particular data point corresponds to genus of that phage. Most phage records are located around the centre, with some distinct groups of *Mycobacterium* and *Staphylococcus* phages outside the centre. Although this could suggest difficulties with distinguishing different genera, it was not the case as the first two principal components retained less than 21\% of dataset variability. Therefore, we assumed a binary representation of phages is reasonable and proceeded with a different method of analysis.

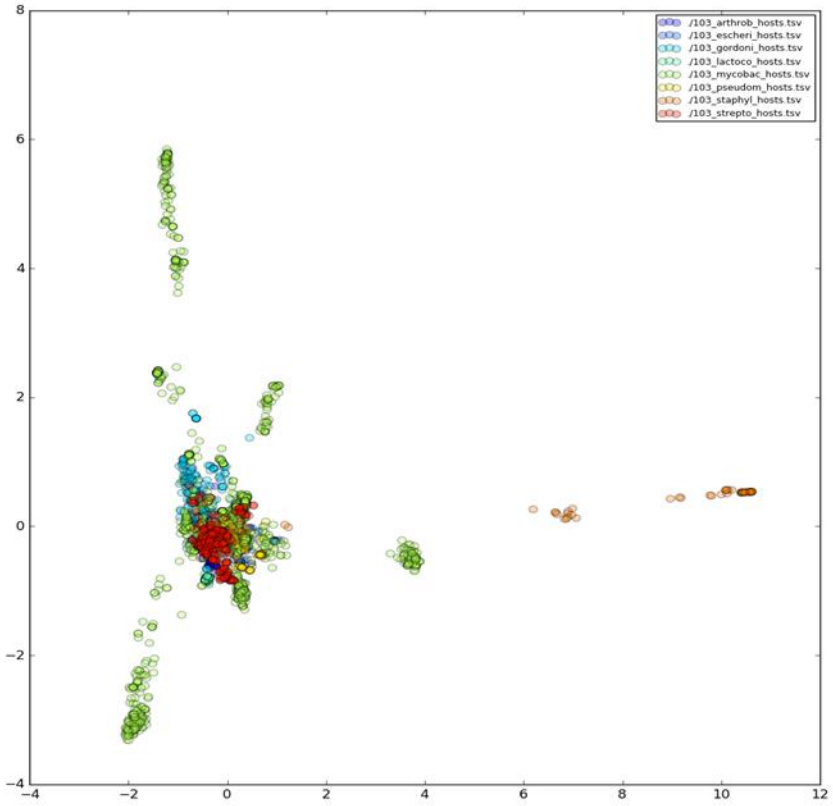

Figure 2: Principal component analysis: first two principal components, PC1 (11.57% of variability) on the x-axis, PC2 (9.19% of variability) on the y-axis

Our training dataset consisted of a matrix with 4723 rows, representing phages and 32281 columns, representing gene clusters. This 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 representable in terms of important clusters rather than in terms of principal components. Because we expected a lot of clusters with small number of genes, our choice for feature selection method was the *emph{Variance Threshold}*. The Variance Threshold method is a simple method that removes

columns with variance under certain threshold. With this technique, all columns in the matrix with ones in more than 99\% of cases or with zeros in more than 99\% of cases were removed. The reduced matrix had 4723 rows and 1965 columns.

For decision tree development, Decision Tree Classifier from python library scikit-learn was used. 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 particular genus. Models were trained with reduced matrix used as features. The ability to infect a particular genus was used as labels. To prevent overfitting of our trees, the parameter *min_impurity_split=0.03* was used. This 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 a model for each of our 50 selected host genera was created and visualized with a python library graphviz. For classification, we expected to have complete sequence of bacteriophage.

Host prediction evaluation

To examine accuracy of our models, all bacteriophages from our test dataset were classified. Test dataset contained 1,202 phage records (Tab.1). 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, the accuracy, sensitivity, specificity and informedness prediction for 50 bacterial genera with the highest number of infecting phages was determined. PHERI best predicted a host for bacteriophages infecting *Leuconostoc*, *Reugenia* and *Helicobacter*. Accuracy, sensitivity and specificity equal to or close to 100%. At

the opposite end of the prediction accuracy spectrum were bacteriophages infecting the genera *Stenotrophomonas*, *Citrobacter* and *Mycobacterium* (Fig. 3). The complete documentation as well as the application itself can be downloaded here (https://hub.docker.com/r/andynet/pheri.). The source code of the tool can be found here (https://github.com/andynet/pheri.).PHERI

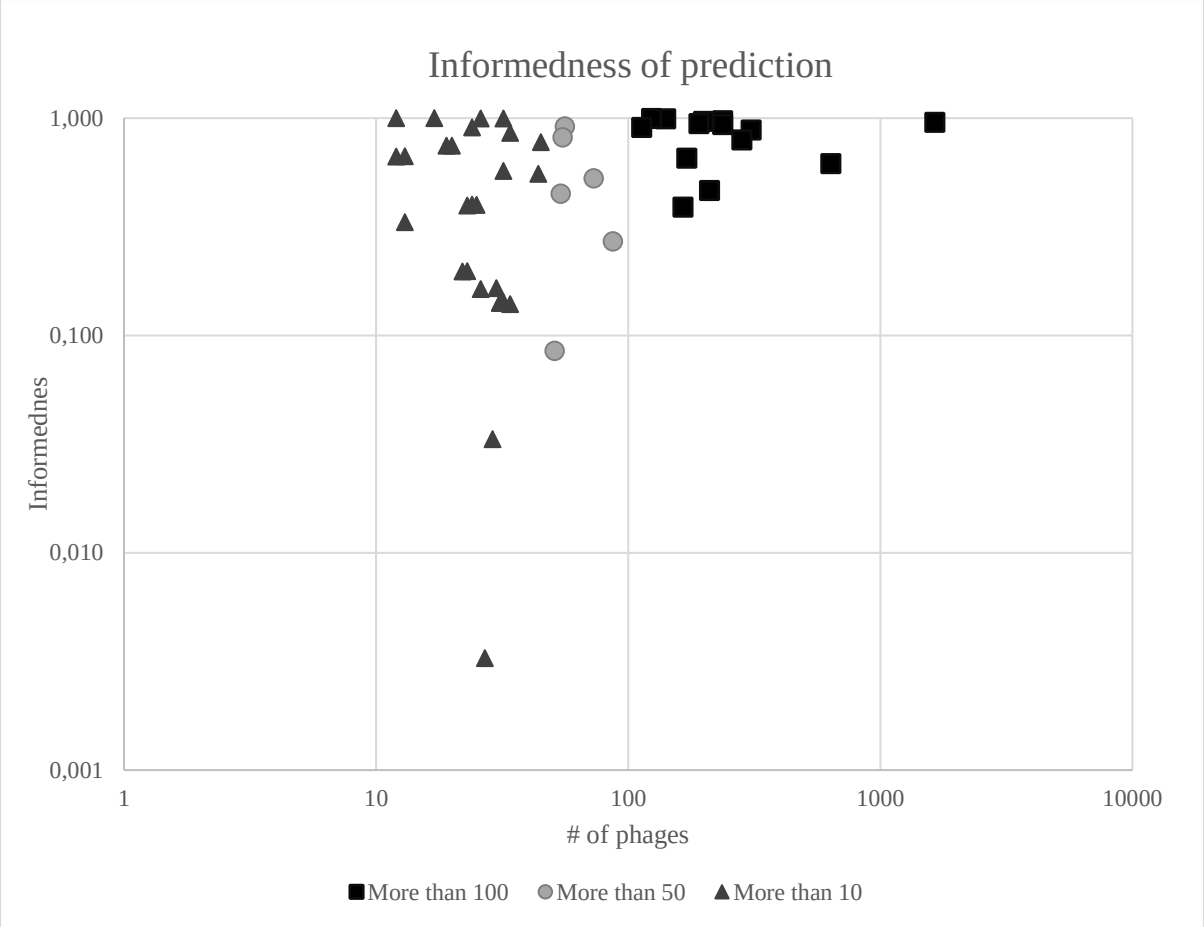

Figure 3: Informedness of PHERI host prediction. Bacterial families were divided into three groups, according to the number of infecting phages in the database, more than 100, more than 50 and more than 10. The value *Informedness* estimates the probability of an informed decision, the closer the values are to one, the more credible the prediction is.

Table 1: The accuracy of the method validated on phage sequences from the testing set

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

Host prediction for new isolated bacteriophages

The functionality of the tool was also verified by determining the host of phages isolated in our lab and were 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. However, for the DevCS-701 phage, PHERI determined different bacterial genus (Tab. 2). According to laboratory tests, bacteriophage Dev-CS701 infects strains from the genus *Cronobacter*, although PHERI predicted *Citrobacter* genus as the most likely candidate. For this reason, the host panel was expanded to include *Citrobacter* strains and specificity was re-established. 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. 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 (Fig. 4). 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 contained six sequences that were classified in five clusters for both decision trees. However, PHERI was able to classified phage only according to the *Citrobacter* decision tree.

Table2. Host prediction for newly isolated and sequenced phages from Slovakia.

Bacteriophage	Closest relative (accession number)	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

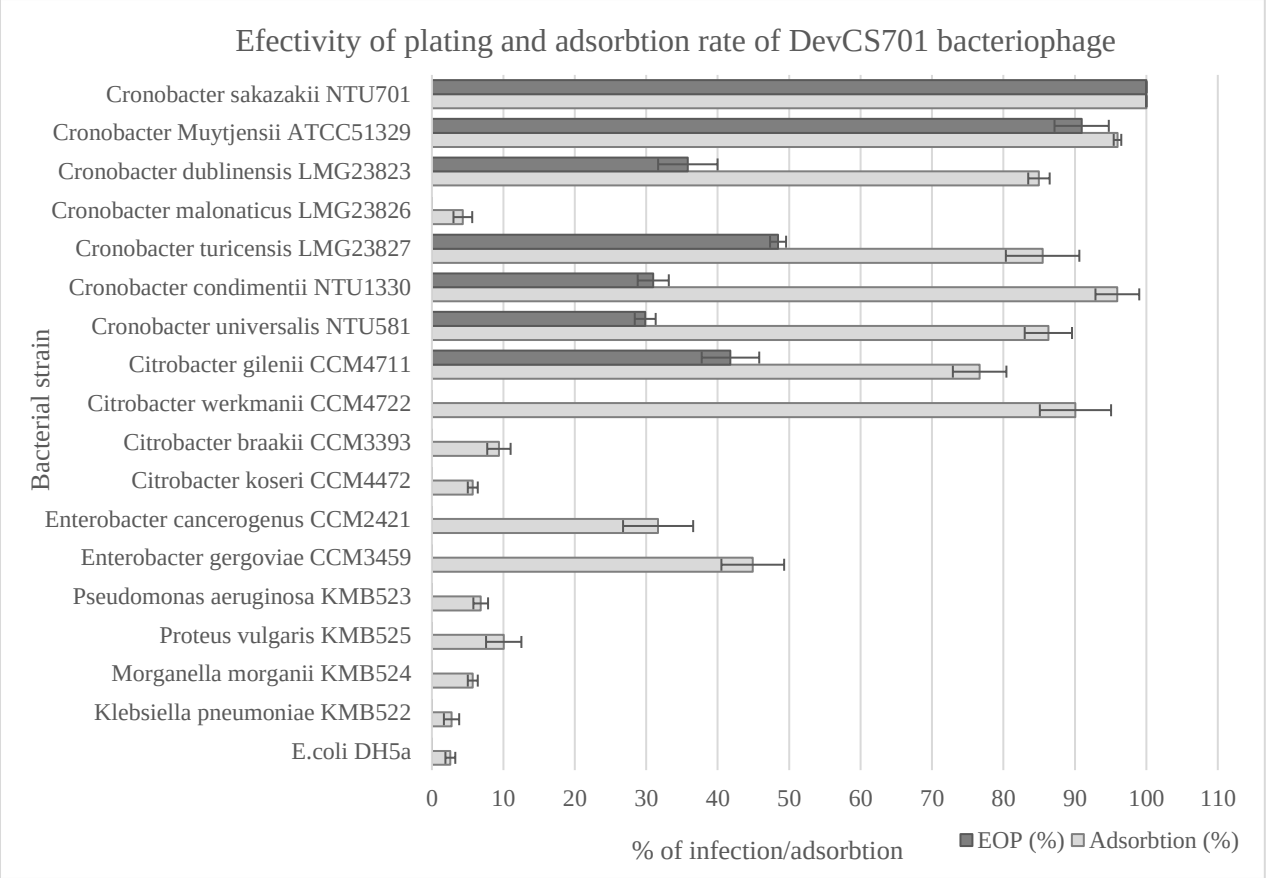

Figure 4: Effectivity of plating and adsorption rate of DevCS701 phage on various hosts.

Discussion

The bacteriophages research could 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 specificity [59][60]. The classic bacteriophage characterization methods were based on phage studies with a known host range and subsequent sequencing, but with the advent of new massively parallel sequencing methods, the procedures often reversed. In addition, 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 to infect foodborne pathogens [61, 62]. However, by exploring new possibilities for phage identification, we obtained metagenomic data containing bacteriophages without a known host. Therefore, we developed a bioinformatics tool based on machine learning algorithms for predicting phage bacterial host genus from the whole-genome sequences, PHERI. A couple of groups have already tested the idea of using a bioinformatic approach to identify phage hosts. One possibility of identifying a host without cultivation has been described by Martínez-Garcia et al.. 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 hyperhalophilic *Nanohaloarchaeota*, included in the so-called 'microbial dark matter'[63] 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 analysed 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)[64]. 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 Willianson 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]. Our tool, bases its prediction on machine learning algorithms. The disadvantage of this pipeline is the dependence on the amount and quality of available data. We used phage whole-genomics records from public databases with known host to create clusters of similar gene sequences that are specific to 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. We assigned a cluster number to all newly obtained genes based on the cluster number of the most similar gene from the BLAST database. Thereafter, vector of ones and zeros was created for each phage representing. This vector was passed to the decision tree model and resulting prediction was saved. Moreover, 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 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 100 phages in the database (Fig.3). The average sensitivity of prediction here was 80%. The prediction was less sensitive for families with more than 50 and more than 10 phages, reaching 56 and 49%, respectively. The data therefore shows that more representatives in the database increase the accuracy of the prediction. This is probably due to the greater number of different host-specific protein sequences that PHERI clustered and incorporated into the decision tree. This reduces the likelihood of incorrect prediction in a case of phage with different mechanism of infection. Improvements in prediction based on machine learning algorithms based on the number of phages in the database have already been described by Chibani et al. 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 had only 17 specific phages in the database. In contrast, in the case of the detection of bacteriophages 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. [65]. PHERI, therefore, constructs a decision tree for a group of genetically-related phages easier and did 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[66], phage vB_SmaS_DLP_5(NC_042082.1) to the *Siphoviridae* family, or phages such as PSH1 (NC_010429.1) to the *Inoviridae* family[67]. Numberless and variable group of phages 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 cases of phages with the ability to infect multiple genera have been described [62][68, 69]. Even in these cases of known cross genera host specificity, only one genus name is found in the database.

We have also used our tool to locate a host of several phages isolated and characterized in our laboratory in the past. The phages had established host specificity for 82 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 (Tab.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 actually recognizes the surface not only of *C. gillenii* but also of *C. werkmanii* (Fig. 4). In addition, a comparison of whole-genome sequences of phages by BLAST showed that Dev-CS701, besides to the closest relative cronophage vB_CsaM_IeB (KX431559.1) [70], 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 possible way in the prediction of phage hosts mainly from metagenomic data. 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

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

1. Ogawara H (2019) Comparison of Antibiotic Resistance Mechanisms in Antibiotic-

Producing and Pathogenic Bacteria. Molecules 24.: https://doi.org/10.3390/molecules24193430

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

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

4. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, et al (2016) Uncovering Earth's virome. Nature 536:425–430

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

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

7. Yu P, Mathieu J, Lu GW, et al (2017) Control of Antibiotic-Resistant Bacteria in Activated Sludge Using Polyvalent Phages in Conjunction with a Production Host. Environmental Science & Technology Letters 4:137–142

8. Simmonds P, Aiewsakun P (2018) Virus classification – where do you draw the line? Archives of Virology 163:2037–2046

9. Yu P, Mathieu J, Li M, et al (2016) Isolation of Polyvalent Bacteriophages by Sequential Multiple-Host Approaches. Appl Environ Microbiol 82:808–815

10. Ye M, Sun M, Huang D, et al (2019) A review of bacteriophage therapy for pathogenic

bacteria inactivation in the soil environment. Environ Int 129:488–496

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

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

13. Anand T, Virmani N, Kumar S, et al (2019) Phage Therapy for treatment of virulent Klebsiella pneumoniae infection in mouse model. J Glob Antimicrob Resist. https://doi.org/10.1016/j.jgar.2019.09.018

14. Dissanayake U, Ukhanova M, Moye ZD, et al (2019) Bacteriophages Reduce Pathogenic Counts in Mice Without Distorting Gut Microbiota. Front Microbiol 10:1984

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

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

17. Carrillo CL, Loc Carrillo C, Atterbury RJ, et al (2005) Bacteriophage Therapy To Reduce Campylobacter jejuni Colonization of Broiler Chickens. Applied and Environmental Microbiology 71:6554–6563

18. Cafora M, Deflorian G, Forti F, et al (2019) Phage therapy against Pseudomonas aeruginosa infections in a cystic fibrosis zebrafish model. Scientific Reports 9

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

20. Jault P, Leclerc T, Jennes S, et al (2019) 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 19:35–45

21. Zhvania P, Hoyle NS, Nadareishvili L, et al (2017) Phage Therapy in a 16-Year-Old Boy with Netherton Syndrome. Frontiers in Medicine 4

22. Hoyle N, Zhvaniya P, Balarjishvili N, et al (2018) Phage therapy against Achromobacter xylosoxidans lung infection in a patient with cystic fibrosis: a case report. Research in Microbiology 169:540–542

23. Law N, Logan C, Yung G, et al (2019) Successful adjunctive use of bacteriophage therapy for treatment of multidrug-resistant Pseudomonas aeruginosa infection in a cystic fibrosis patient. Infection 47:665–668

24. Aslam S, Courtwright AM, Koval C, et al (2019) Early clinical experience of bacteriophage therapy in 3 lung transplant recipients. Am J Transplant 19:2631–2639

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

26. Hyman P, Abedon ST (2010) Bacteriophage Host Range and Bacterial Resistance. Advances in Applied Microbiology 217–248

27. Tu J, Park T, Morado DR, et al (2017) Dual host specificity of phage SP6 is facilitated by tailspike rotation. Virology 507:206–215

28. Hutinet G, Kot W, Cui L, et al (2019) 7-Deazaguanine modifications protect phage DNA

from host restriction systems. Nat Commun 10:5442

29. Furi L, Crawford LA, Rangel-Pineros G, et al (2019) Methylation Warfare: Interaction of Pneumococcal Bacteriophages with Their Host. J Bacteriol 201.: https://doi.org/10.1128/JB.00370-19

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

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

32. Chen B, Akusobi C, Fang X, Salmond GPC (2017) 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 8

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

34. Roux S, Enault F, Hurwitz BL, Sullivan MB (2015) VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985

35. Williamson SJ, Allen LZ, Lorenzi HA, et al (2012) Metagenomic exploration of viruses throughout the Indian Ocean. PLoS One 7:e42047

36. Villarroel J, Kleinheinz KA, Jurtz VI, et al (2016) HostPhinder: A Phage Host Prediction Tool. Viruses 8.: https://doi.org/10.3390/v8050116

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

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

39. Chibani CM, Meinecke F, Farr A, et al ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks

40. Lopes A, Tavares P, Petit M-A, et al (2014) Automated classification of tailed bacteriophages according to their neck organization. BMC Genomics 15:1027

41. Enright AJ (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30:1575–1584

42. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410

43. Benson DA, Cavanaugh M, Clark K, et al (2018) GenBank. Nucleic Acids Res 46:D41–D47

44. Cock PJA, Antao T, Chang JT, et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423

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

46. Cosma CL, Sherman DR, Ramakrishnan L (2003) The secret lives of the pathogenic mycobacteria. Annu Rev Microbiol 57:641–676

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

48. Hyatt D, Chen G-L, Locascio PF, et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119

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

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

51. Haft DH, Loftus BJ, Richardson DL, et al (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29:41–43

52. Altschul S (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410

53. Tristão Ramos RJ, de Azevedo Martins AC, da Silva Delgado G, et al (2017) CrocoBLAST: Running BLAST efficiently in the age of next-generation sequencing. Bioinformatics 33:3648–3651

54. Heringa J (2004) Needleman-Wunsch Algorithm. In: Dictionary of Bioinformatics and Computational Biology

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

56. Oliphant TE (2007) Python for Scientific Computing. Comput Sci Eng 9:10–20

57. Boulanger P (2009) Purification of bacteriophages and SDS-PAGE analysis of phage structural proteins from ghost particles. Methods Mol Biol 502:227–238

58. Garreta R, Moncecchi G (2013) Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd

59. Mahony J, McAuliffe O, Paul Ross R, van Sinderen D (2011) Bacteriophages as biocontrol

agents of food pathogens. Current Opinion in Biotechnology 22:157–163

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

61. Kajsík M, Bugala J, Kadličeková V, et al (2019) Characterization of Dev-CD-23823 and Dev-CT57, new Autographivirinae bacteriophages infecting Cronobacter spp. Arch Virol 164:1383–1391

62. Kajsík M, Oslanecová L, Szemes T, et al (2014) Characterization and genome sequence of Dev2, a new T7-like bacteriophage infecting Cronobacter turicensis. Arch Virol 159:3013–3019

63. Martínez-García M, Santos F, Moreno-Paz M, et al (2014) Unveiling viral-host interactions within the "microbial dark matter." Nat Commun 5:4542

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

65. Kot W, Hansen LH, Neve H, et al (2014) Sequence and comparative analysis of Leuconostoc dairy bacteriophages. Int J Food Microbiol 176:29–37

66. Fan H, Huang Y, Mi Z, et al (2012) Complete Genome Sequence of IME13, a Stenotrophomonas maltophilia bacteriophage with large burst size and unique plaque polymorphism. J Virol 86:11392–11393

67. Liu J, Liu Q, Shen P, Huang Y-P (2012) Isolation and characterization of a novel filamentous phage from Stenotrophomonas maltophilia. Archives of Virology 157:1643–

68. McCutcheon J, Peters D, Dennis J (2018) Identification and Characterization of Type IV Pili as the Cellular Receptor of Broad Host Range Stenotrophomonas maltophilia Bacteriophages DLP1 and DLP2. Viruses 10:338

69. Peters DL, Lynch KH, Stothard P, Dennis JJ (2015) The isolation and characterization of two Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. BMC Genomics 16

70. Endersen L, Buttimer C, Nevin E, et al (2017) 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 253:1–11