OXFORD

Briefings in Bioinformatics, 2024, **25(3)**, bbae117
https://doi.org/10.1093/bib/bbae117
Review

Advances in phage–host interaction prediction: *in silico* method enhances the development of phage therapies

Wanchun Nie†, Tianyi Qiu†, Yiwen Wei, Hao Ding, Zhixiang Guo and Jingxuan Qiu

Corresponding author: J. Qiu, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 334, Jungong Road, Yangpu District, Shanghai 200093, China. Tel.:/Fax: 021-55270818; E-mail: jxqiu@usst.edu.cn
†Wanchun Nie and Tianyi Qiu contributed equally to this work.

Abstract

Phages can specifically recognize and kill bacteria, which lead to important application value of bacteriophage in bacterial identification and typing, livestock aquaculture and treatment of human bacterial infection. Considering the variety of human-infected bacteria and the continuous discovery of numerous pathogenic bacteria, screening suitable therapeutic phages that are capable of infecting pathogens from massive phage databases has been a principal step in phage therapy design. Experimental methods to identify phage–host interaction (PHI) are time-consuming and expensive; high-throughput computational method to predict PHI is therefore a potential substitute. Here, we systemically review bioinformatic methods for predicting PHI, introduce reference databases and *in silico* models applied in these methods and highlight the strengths and challenges of current tools. Finally, we discuss the application scope and future research direction of computational prediction methods, which contribute to the performance improvement of prediction models and the development of personalized phage therapy.

Keywords: phage–host interaction; bacteriophage; phage therapy; in silico model

INTRODUCTION

As the most abundant organism on the earth, bacteriophage was first discovered by Frederick William Twort and Félix d'Herelle in the early 1900s [1, 2], which can be found everywhere with their host bacteria. Phages have been found with the potential for clinical detection [3], design of vaccines [4], food preservation [5] and wastewater treatment [6], which rely on the bactericidal action of phages. In addition, with the rapid emergence of drug-resistant bacteria caused by the overconsumption of antimicrobials, phage therapy is now considered with therapeutic effects for the infection of drug-resistant bacteria [7]. Therefore, identification of phage–host interaction (PHI) is essential for investigation of phage and further usage of phage therapy.

Currently, the experimental method is still the golden standard to recognize PHI, which includes plaque assays, liquid assays, viral tagging, microfluidic PCR, phageFISH, single-cell sequencing and Hi-C sequencing [8]. Plaque assays is the most common method among these methods, which identify PHI by 'plaque'. This method cultured a single bacterial strain on the agar layer, and phage was spotted on the layer. Plaque on the agar layer after the cultivation signifies that the bacterial strain was infected by the phage, and the conclusion that the cultured bacterium is the host of the phage will be proved.

Phages can specifically recognize the receptors on the surface of bacterial cells, and then the genome of the bacteriophage will be injected into the bacterial cell and be replicated with the bacterial genome [9]. To evade phage infection, bacteria have correspondingly evolved various mechanisms, which can be classified according to the different stages of PHI (Figure 1). At the first stage of PHI, bacteria have evolved the first line of defense to block phage adsorption with the receptor on the bacterial cells [10], and phage correspondingly developed new mechanisms to escape the bacterial defense, such as recognizing new bacterial receptor. At the second line of defense, bacteria have developed with superinfection exclusion to defend phage infection by preventing genetic material from entering bacterial cells [11]. In addition, bacteria have numerous intracellular immune systems to defend after phages injecting genome sequences into bacterial cell, such as restriction-modification system [12] and CRISPR-Cas system [13] and Toxin-antitoxin system [14]. As a response, bacteria evolved

Wanchun Nie is currently a master student at the School of Health Science and Engineering, University of Shanghai for Science and Technology. Her research interests are bioinformatics and machine learning.
Tianyi Qiu is a professor at the Institute of Clinical Science, Zhongshan Hospital, Fudan University. His research interests are bioinformatics and immunoinformatics.
Yiwen Wei is currently a master student at School of Health Science and Engineering, University of Shanghai for Science and Technology. Her research interests are bioinformatics and vaccine design.
Hao Ding is currently a master student at Fudan University. His research interests are bioinformatics and computer science.
Zhixiang Guo is currently a master student at School of Health Science and Engineering, University of Shanghai for Science and Technology. Her research interest is Modification of natural active substances.
Jingxuan Qiu is an Associate Professor at the School of Health Science and Engineering, University of Shanghai for Science and Technology. Her research interests are bioinformatics and pathogen biology.
Received: September 10, 2023. **Revised:** January 15, 2024. **Accepted:** March 2, 2024

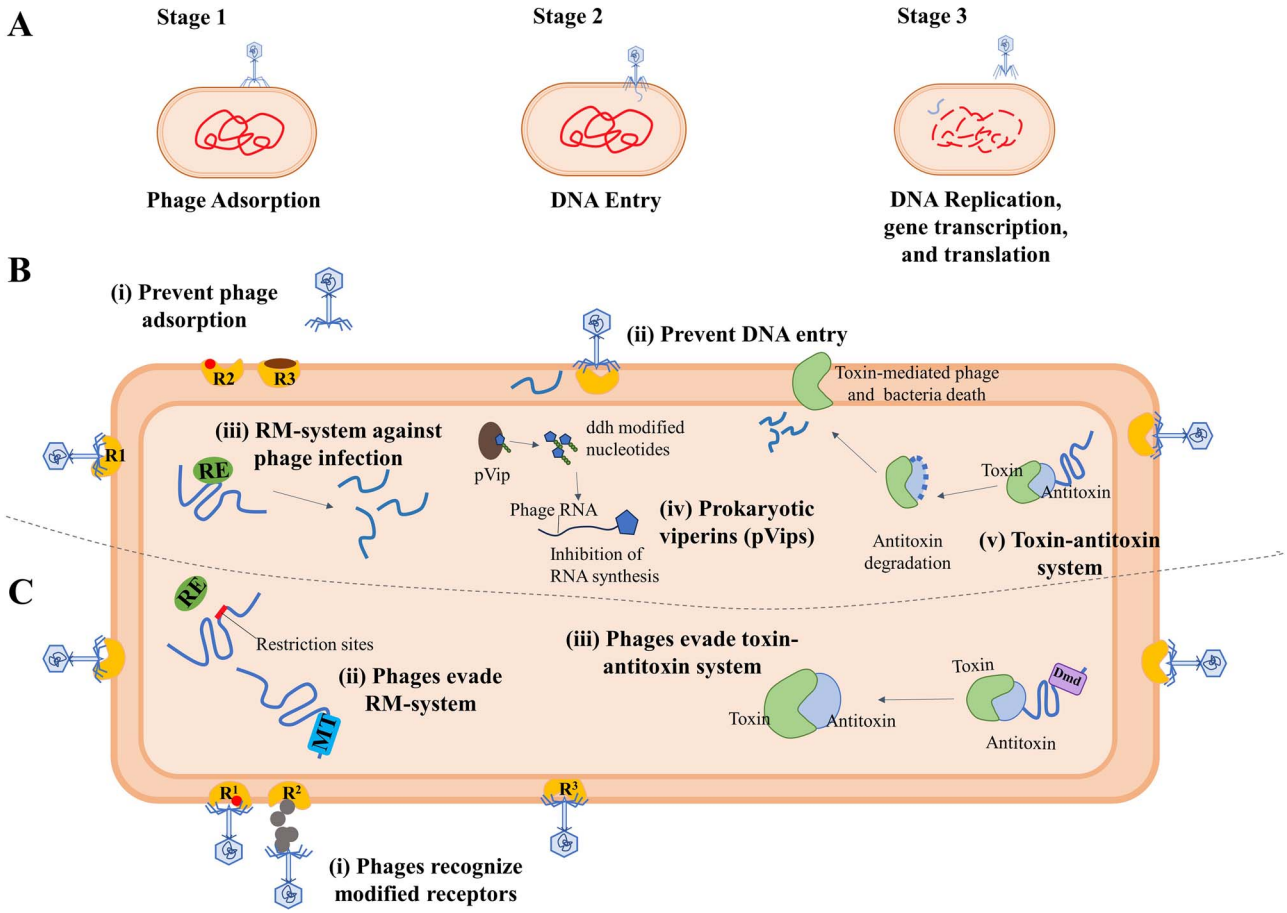

Figure 1. The arms-race between phage and host. (**A**) The different stages of phage infection. (**B**) The mechanism of bacterial defenses against phage infection during the arms-race. (i–v) Representing five different steps for bacteria to defend against phage infection, which include (i) prevent phage adsorption (contains three strategies: R1 represents normal phage receptor, R2 represents receptor mutant and R3 represents protein modified receptor). (ii) Prevent phage DNA entry into the bacterial cell. (iii) Restricted modification systems (RM-system) against phage infection. RE is a restriction endonuclease and MT is a methyltransferase. RM-systems can cleave phage DNA upon recognition of specific sequence motifs. (iv) Prokaryotic viperins (pVips). pVips can produce RNA chain terminator molecule to inhibit phage RNA synthesis. (v) Toxin-antitoxin system. Toxin-antitoxin systems can be activated upon phage infection and lead to cell death or growth arrest. (**C**) The mechanism of phages evasion from bacterial defenses during the arms-race. (i–iii) Representing three different steps for phages evade from bacterial defenses, including (i) bacteriophages counteract bacterial adsorption inhibition (contains three strategies: R^1 represents phage recognizes mutant receptor, R^2 represents phage degrades the receptor's polysaccharide capsule and R^3 represents phage recognizes new receptor of the bacteria). (ii) Phages evade RM-system. Phages evade the RM systems by deleting the restriction sites and using the host's methyltransferase to modify its DNA. (iii) Phages evade toxin-antitoxin system. Phage encodes antitoxin factor Dmd to block toxin-antitoxin system.

with two main strategies including inactivation of the immune systems by encoding protein inhibitors and bacteriophage gene modifications or mutations [15]. The coevolution of bacterial defense against phage infection and phages evading bacterial defense leads to the evolutionary arms-race [16], and phages adapt their composition of the genome to their host bacterial genome during the co-development of phages and their hosts. The coevolution of phages and bacteria brings the phage–host genomic association, which has been regarded as the signal of PHI [8]. Accordingly, many computational phage–host prediction models have been constructed basing on the signal of genomic association between phages and bacteria.

Computational methods apply machine learning algorithms to deal with signals indicating phage–bacteria interaction to predict the relationship between query phage and bacteria, which can be broadly classified into classification and regression. Current computational methods have been capable of predicting hosts for query phage on the genus or even species level. Although the predicted result cannot represent exact PHI, it shows a great perspective in phage therapy.

METHODS FOR PHI PREDICTION

Bacteriophages can specifically recognize the receptors including receptor-binding protein (RBP), polysaccharide and lipopolysaccharide on the surface of the host bacterial cell, and then integrate genome segments into the genome of their bacterial host. Based on the phage infection mechanism, nucleotide similarity and amino acid similarity between phages and bacterial genome were discovered, which were driven by specific genetic fragments such as host-encoded clustered regularly interspaced short palindromic repeats (CRISPR) spacers [17], exact matches [8] and tRNAs [18]. Taking the similarity of genome sequence and gene character as signals, computational methods were therefore proposed to predict the interaction of phage and bacteria, which have been developing at a rapid speed. The computational methods can be

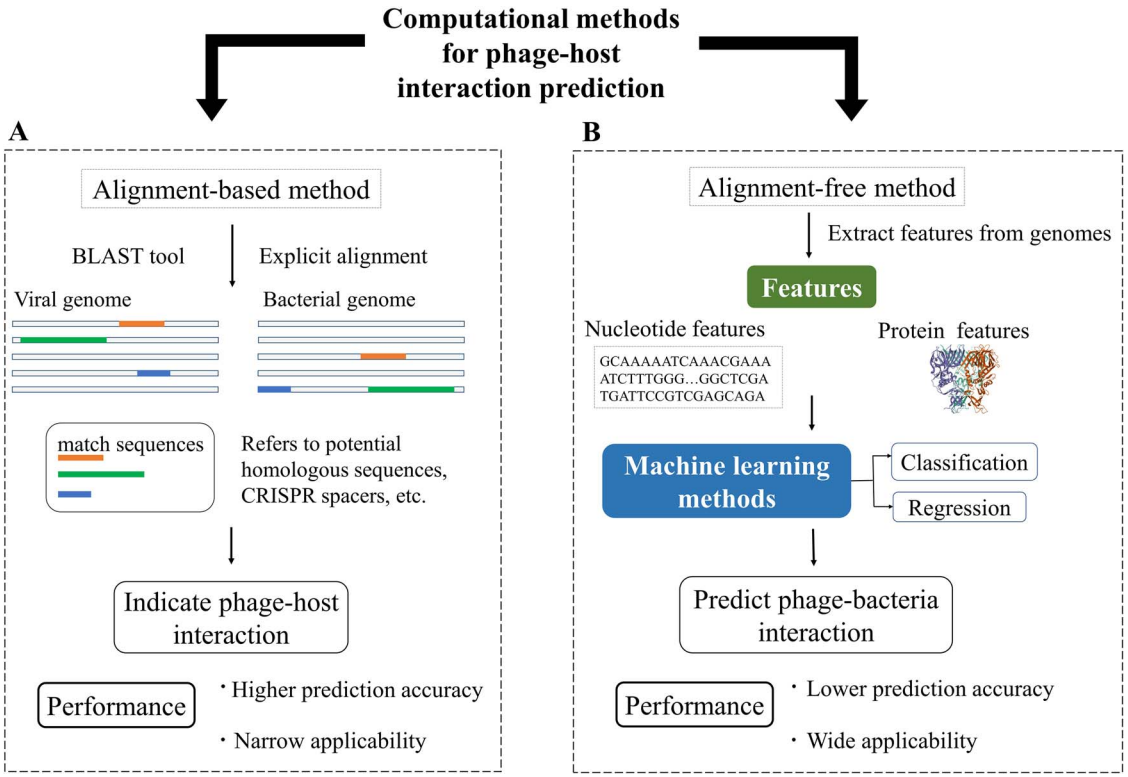

Figure 2. Mechanization of computational methods for PHI prediction. (**A**) Alignment-based methods achieve interaction identification by explicit alignment of viral and bacterial whole-genome sequences and acquire matched sequences that indicate PHI. (**B**) Alignment-free methods compare nucleotide features and/or protein features extracted from viral and bacterial genomes, and predict PHI by optimized machine learning methods. Alignment-based methods show higher prediction accuracy, whereas alignment-free methods have wider applicability.

divided into two categories: (1) alignment-based method and (2) alignment-free method. The mechanism of these two types of computational methods is expounded in Figure 2.

Alignment-based method

Alignment-based methods predict PHI with explicit alignment of viral and bacterial genome sequences to figure out the potential shared regions between phages and bacteria.

Methods based on nucleotide sequence similarity

Alignment-based method was first associated with the Basic Local Alignment Search Tool [19] to achieve the alignment of phages and bacterial genomes. This BLAST-based approach analyzes the sequence similarity to identify homology sequences and/or shared genes between phages and bacteria [8]. By inputting the genome sequence of query phage, BLAST can predict the bacterial host for query phage from the publicly available reference databases by analyzing the *E*-value, bit score, match length and number of mismatches between the genome sequences of phage and bacteria. However, the inference of phage–host relationships through BLAST-based methods is limited by the comprehensiveness and completeness of the used databases because of the high reliance on the database. In the situation that the alignment result shows little genome similarity between query phages and bacteria collected in the reference database, BLAST cannot predict the host for query phage at all. In addition, some phages show cross-reaction [20], the BLAST-based method ranked score for multiple bacteria with user-defined criteria and can only predict one possible host with the highest score for query phage.

Therefore, CRISPR-based method was proposed for more precise alignment. In recent research, it was found that bacteria can employ CRISPRs along with associated Cas genes to resist the invasion of bacteriophages [21]. For some prokaryotic organisms, short sequences of phage and plasmid origin which are called CRISPR spacers will be integrated into the CRISPR array of the bacterial genome. These CRISPR spacers are transcribed and processed into CRISPR RNAs, which combine with Cas genes and act as guides for site-specific cleavages of the invader's genome [22]. During this situation, consistent sequences were found in both genomes of the bacteriophage and its host; an array of nucleotide sequences of the bacteriophage will be incorporated into a CRISPR spacer, and be transmitted to the progeny bacteria. Thus, alignment between host-encoded CRISPR spacers and viral genome can be effective for PHI prediction. With developing research on the CRISPR system, some bioinformatic tools such as PILER [23] and CRISPRDected [24] were widely used to detect CRISPR spacers and CRISPR arrays from bacterial genomes. CRISPR-Cas immune system was originally used to identify viral and bacterial genomes from various metagenomes, including the human microbiome [25], marine environment [26], glacial ice and soil [27]. These researches utilized host-encoded CRISPR spacers as gene probes to identify phage genomes, which distinguish phages and bacteria genomes among metagenomes without constructing any bioinformatic prediction tool. After that, the first CRISPR-based tool to predict host for uncharacterized phages was created by Moira [28], it compared genetic similarity between viral genome and CRISPR spacers with BLASTN to predict host as the bacterium with spacers targeting most regions of phage genome. This method shows good prediction performance but low sensitivity because of strict filtering criteria, for example, the small number of mismatches. To

improve this weakness of CRISPR-based methods, SpacePHARER [29] was constructed as a sensitive tool to predict phage–host relationship. It considered multiple hits for each query CRISPR spacers in a bacterial genome to improve prediction accuracy and built a negative viral data set for the optimistic false discovery rate to improve sensitivity. In addition, the prediction performance depends heavily on the maximum number of mismatches that are allowed between CRISPR spacers and phage genome, and there are only 40% of bacteria encoding CRISPR system to date [30], which leads to the fact that CRISPR-based method cannot be effectively applied to a great number of bacteria.

Methods based on viral marker genes

Some other alignment-free methods regard similar genes between phages and bacteria as the indication of PHI, which is applicable for phages at least sharing one marker gene with a known phage reference. For example, Random Forest Assignment of Hosts (RaFAH) constructed a Random Forest classifier which is applied for protein domain data, identifying the combination of genes between phage and bacteria that are indicative of phage–host associations [31]. Based on the genomes of uncultured viruses derived from eight different biomes, RaFAH compared the query protein to a custom database of HMM profiles obtained from isolated phages and uncultivated phages with a high-confidence host prediction, and the HMM profiles identified for the query phage were inputted to Random Forest classifier to output a prediction score for each candidate host from phylum to genus.

Alignment-based approaches predict bacteriophage–host interaction mainly depends on the analysis of homologous genome sequences and shared genes between phage and host. However, newly discovered phages and bacteria increased dramatically, many of which have not yet been detected with homologous genome sequences and shared genes. Thus, the alignment-based method can only be applied in a narrow range. Furthermore, alignment-based approaches often show high accuracy but low recall when using strict criteria, that is, the percentage of the phage–bacteria interaction predicted to be correct is high, whereas the percentage of correct phage–host pair is low compared with the total number of input phages.

Alignment-free method

Instead of explicit alignment between viral and bacterial genomes, another group of methods for PHI prediction based on the similarity of sequence composition also developed speedily. Phages adapt their composition of the genome to their host bacterial genome since that the expression of viral genes depends on the translation mechanism of host bacteria [31]. This process may result from the exchange of hereditary substances, the coevolution of regulatory sequences, the adaption of the codon usage to the tRNA pool available in the host cell and/or the evasion of the host defense systems. The similar composition between phages and bacterial genomes makes it possible to predict PHI with the alignment-free method, which can assess the interaction between phages and bacteria without discovering homologous genome sequences. Instead of predicting the interaction between cultured phage and bacteria with an exact match of homologous genomes, alignment-free method is also available to predict potential hosts for dramatically increased uncultured phages.

Methods based on viral and bacterial co-abundance

Finding that genomes of bacteriophages and bacteria are consistent in time and space, abundance profiling method was proposed as the first alignment-free method for PHI prediction, which can be applied to infer PHI not only for temperate phages that integrate genome fragments to their host bacterial genome but lytic phages that depend on their host for survival. Current researches have certified the abundance consistent of metagenome between the bacteriophages and bacteria in the human gut [32] and the environment of the ocean [33]. To confirm the coherence of viral and host bacterial abundance, Edwards [8] applied FOCUS [34] and MEGABLAST [35] to identify the phage and bacteria present, respectively, and verified the usability by analyzing abundance profiles of each phage and bacterial genome with Pearson correlation. Nevertheless, abundance consistent of phage and bacteria cannot entirely represent the interaction between phage and bacteria; lag of abundance covariation between phage and bacteria leads to the fact that signals of numerous PHIs cannot be reflected on correlation of viral and bacterial abundance based on current sampling time of metagenome. As a result, the sensitivity of the abundance-based method to detect PHI leads to a low prediction correct rate.

Methods based on nucleotide composition

Another alignment-free method concentrates on the composition of nucleotide, as phages and their hosts often share similar patterns in codon usage or short nucleotide words. Among these, the similarity of oligonucleotide frequency, which is also defined as the similarity of k-mer frequency (k-mer represents segments of DNA with length of k), was widely used as the signal for bacteriophage–host interaction prediction. Oligonucleotide was initially used as a measurement of genetics in Woese and Fox's study [36]; they analyzed the oligonucleotide in the 16S ribosomal RNA from organisms of three kingdoms and found the dissimilarity of oligonucleotide among three kingdoms to demonstrate three aboriginal lines of descent. Researches based on k-mer analysis have referred to genome assembly [37], classification of metagenomes [38], annotation of metagenomes [39] and phage–host identification [40].

Similarity of k-mer composition was first applied as the signal for PHI prediction in Julia's study [40], a phage host prediction tool HostPhinder was constructed based on the genetic similarity that defined as the number of co-occurring k-mer between query phage and phages in the reference database. In view of the fact that there are no common 16S ribosomal RNA genes among all phages, HostPhinder analyzed the oligonucleotide frequency of whole-genome sequences of the phages. It first classified the genome of bacteriophage into different clusters by the number of 16-mers in the genome sequences, and then assigned the query phages to a cluster. The host of the query phage was predicted as the host bacteria of the phage which was classified in the same cluster. However, the prediction accuracy of HostPhinder depends on the breadth of the constructed host database, that is, the reference database of HostPhinder needs to be uploaded constantly. 2016

Different from HostPhinder which compared the k-mer frequency of query phages to the phage reference database, some other methods such as the oligonucleotide-based method in Edward's study [8], VirHostMatcher (VHM) [41] and Prokaryotic virus Host Prediction (PHP) [42] compared the similarity of k-mer frequency between query viral genomes and genomes from a host reference database. Edwards calculated the Euclidean distance of viral and bacterial k-mer usage profiles of length from 3 to 8 bp and validated 4 bp as the best length. Similar studies for PHI prediction that prior to VHM almost utilized Euclidean distance as the only measure of k-mer similarity [8, 43], VHM therefore applied 11 measurements to calculate the dissimilarity of k-mer and verified that d_2^* had the strongest potential to predict the

bacteriophage–host interaction, with a corresponding optimal length of k-mer that screened as 6. PHP then screened the Gaussian model as the optimal model to score for similarity of k-mer frequency between query prokaryotic virus and every bacterium in the reference database, and bacteria with the highest score was predicted as the host. Compared with VHM, PHP shows better prediction accuracy from kingdom to genus level [42].

Methods for PHI prediction mentioned above, however, show worse performance for phage and bacteria with genomes of short contigs, whereas short contigs widely exist in metagenomic data because of nonuniform read coverage of metagenome samples [44]. 'Who Is the Host' (WIsH) [45] achieved better prediction performance for viral and bacterial genomes consisting of contigs as short as 3 kbp. It constructed a homogeneous Markov model of order 8 for each potential host genome basing on similar k-mer-based comparison, and calculated the likelihood of viral contigs with each model, the one whose model yields the highest likelihood is considered as the predicted host for query phage.

It was found that previous studies show low prediction accuracy on the species level, LMFH-VH [46] and ILMF-VH [47] then constructed heterogeneous network of viruses and hosts for PHI prediction. LMFH-VH first constructed a virus similarity network based on k-mer similarity and host similarity network with Gaussian Interaction Profile kernel similarity, and connected these two networks with phage–host association network to establish neighborhood regularized logistic matrix factorization model. The model can calculate the score between query phage and every candidate host, and predict the host for query phage as the top-ranked host. Compared with LMFH-VH, ILMF-VH applied similarity network fusion to construct host network, and constructed heterogeneous network by integrating the virus network, host network and virus–host association network. However, these network-based ignored the sparsity and unconnectedness in the phage–host heterogeneous information network, in order to address the limitation, Wang proposed GERMAN-PHI to predict PHI, which learned representation of phage and bacteria with graph attention neural network and restricted phage–host association matrix by neural inductive matrix completion [48].

In addition, the above nucleotide-based alignment-free methods predict PHI by analyzing the nucleotide similarity of viral and bacterial whole-genome sequences. However, genome similarity usually exists in the local sequences of phages and bacteria because phages can infect host bacteria by integrating genome segments into bacterial genome. In this regard, PB-LKS screened the most similar genome segments between phages and bacteria and analyzed the local k-mer similarity [49]. It is claimed that PB-LKS outperformed other state-of-the-art alignment-free methods and are comparable to the outperformed alignment-based BLAST methods, which shows that the local nucleotide similarity may be a more significant signal of PHI compared with the nucleotide similarity based on whole-genome sequences.

Methods based on protein characteristics

Instead of concentrating on nucleotide composition, alignment-free methods can also analyze the similarity of amino composition and/or protein structure between phages and bacteria for correlation prediction. Leite's study [50] proposed a method to predict phage–bacteria interaction based on the protein–protein interaction of their genomes. It extracted features including domain–domain interaction scores and protein primary structure information, and constructed a prediction model with artificial neural networks. However, the lack of independent test in Leite's study leads to probable overfitting problem. Moreover, negative samples in Leite's study were randomly selected, which might cause unstable model result. Compared with Leite's study, PredPHI [51] selected negative samples with high quality based on the K-means clustering method and achieved better robustness. PredPHI identified PHI by analyzing sequences and structural information of protein sequences encoded by bacteriophage and its host, and constructed prediction model with deep convolutional neural network. This method improves prediction performance by applying complex architecture of deep learning models, yet it is uninterpretable with black boxes and is hard to understand for users.

In contrast, tools based on traditional machine learning methods are more interpretable. As a phage host prediction tool, PHERI [52] predicted host for query phages with the assumption that phages infecting the same host share similar protein sequences. It created a model with Decision Tree Classifier for phages that infect bacteria on a specific genus basing on phages' annotated protein sequences clustered by TRIBE-MCL [53], which were used to testify whether the query phages were able to infect bacteria on the specific genus. This method not only shows great prediction accuracy in the era of big data sets, but also shows great potential for highlighting infection-related genes with unknown functions. In Boeckaerts's study [54], they employed random forest as optimal machine learning methods to predict bacteriophage hosts based on sequences of annotated RBP. RBP can be considered as a signal for PHI with the basic that RBP can recognize specific bacterial receptors on the bacterial cell surface and is regarded as the determinant for the specificity of phage infection [55]. This RBP-based method took nucleotide sequences along with protein sequence and structure data as comprehensive features basing on bacterial and viral sequences annotated as RBP protein rather than whole-genome sequences, and assigned query phage to the class of a specific bacterial host species in the constructed RBP database. This method combined protein characters and nucleotide characters as comprehensive considerations to identify phage–host relationship. Similarly, HostG [56] and PHIAF [57] also analyzed both nucleotide and amino acids to recognize PHI. As a graph convolution network (GCN)-based semi-supervised learning approach, HostG constructed a network of viruses and host by analyzing virus–virus protein similarity and virus–host DNA sequence similarity, and the prediction of HostG mainly depended on the constructed knowledge graph of phages and host, of which the node feature and edge connection were on the basic of the interaction of phage and host along with the viral and host sequences. For PHIAF, it constructed a data augmentation module with generative adversarial network (GAN) to overcome the scarcity of PHIs, and achieved prediction by the CNN model with fused DNA and protein features of the augmented data. All of the above methods relied on constructing handcraft genomic and protein features; however, a recent study proposed to generate dense vector encodings of RBPs with protein language model and verified better performance than other handcraft feature-based methods [58].

DATABASE OF COMPUTATIONAL APPROACHES FOR BACTERIOPHAGE–HOST INTERACTION PREDICTION AND ASSISTANCE TOOLS

With the rapid development of the next-generation genome sequencing technology, computational approaches were therefore an alternative for bacteriophage–host interaction prediction,

Table 1: Databases of PHI

Database	Website	Data volume (up to 21 July 2023)	Character
NCBI RefSeq genomes	https://www.ncbi.nlm.nih.gov	4194 interactions between phages and prokaryotic organisms	Comprehensive database
Microbe Versus Phage	http://mvp.medgenius.info	26 572 interactions between 18 608 viral clusters and 9245 prokaryotes	Comprehensive database
Viral Host Range	https://viralhostrangedb.pasteur.cloud	171 701 interactions between 776 viruses and 2041 hosts.	Comprehensive database
PhagesDB	https://phagesdb.org	4487 interactions between 4487 actinobacteriophages and 114 hosts	Database specifically collected virus–host interaction for actinobacteriophage

Table 2: Typical genome databases applied in phage–host prediction

Phage–host prediction tool	Reference database	Website	Data feature	Data analyzation
HostPhinder [48]	NCBI RefSeq genomes	https://www.ncbi.nlm.nih.gov	Viral genome	K-mer frequency
VHM [49]	NCBI RefSeq genomes		Viral and bacterial genome	K-mer frequency
PHP [50]	NCBI RefSeq genomes		Viral and bacterial genome	K-mer frequency
LMFH-VH [55]	NCBI RefSeq genomes		Viral and bacterial genome	K-mer frequency
ILMF-VH [56]	NCBI RefSeq genomes		Viral and bacterial genome	K-mer frequency
GCN [65]	NCBI RefSeq genomes		Viral genome	Genome sequence similarity
WIsH [54]	KEGG	https://www.kegg.jp	Prokaryotic genome	K-mer frequency
PHERI [59]	ViralZone	https://viralzone.expasy.org	Viral genome	Coding-DNA sequences
PredPHI [58]	NCBI RefSeq genomes		Viral and bacterial genome	K-mer frequency
	GenBank	https://www.ncbi.nlm.nih.gov/genbank	Viral and bacterial genome	K-mer frequency
RaFAH [39]	NCBI RefSeq genomes		Viral genome	Genome sequence similarity
	Pfam	http://pfam.xfam.org	Viral protein families	Protein domain
GSPHI [71]	NCBI RefSeq genomes		Viral genome	Genome sequence similarity
	UniProtKB	https://www.uniprot.org	Viral and bacterial protein	RBP
	MillardLab	https://millardlab.org	Viral and bacterial protein	RBP
Leite's study [57]	GenBank		Viral and bacterial genome	Coding-DNA sequences
	Pfam		Viral and bacterial protein families	Protein domain
RBP [61]	UniProtKB		Viral protein sequences	RBP
	MillardLab		Viral protein sequences	RBP

which highly rely on the reference database. Predicting bacteriophage–host interaction with computational methods requires to equip with the information of bacteriophage–host interaction and genome data of phage and bacteria, correlative databases are listed in Tables 1 and 2. Comprehensive bacteriophage–host interaction information was commonly collected in the National Center for Biotechnology Information (NCBI) RefSeq [59], which integrates host information on the genus or species level for numerous phages. Recently, Microbe Versus Phage database [60], Viral Host Range database [61] and Virus–Host database [62] also integrated host range data for bacteriophage. Some other databases receive information for a certain class of bacteria, such as PhagesDB databases [63], which specifically collect bacteriophage–host interaction for actinobacteriophage.

Equipped with information on PHI, genome sequences of related phage and bacteria are required for further analysis. Commonly used genome databases such as the NCBI RefSeq and GenBank database collected genome sequences for numerous phages and bacteria, which were widely applied in computational phage–host prediction tools for genome sequences to achieve nucleotide and protein sequences analysis. What is more, the KEGG database [64] also performed as a reference database to provide prokaryotic genome sequences in WIsH [45]. For some bioinformatic tools of phage–host prediction that are based on the analysis of protein sequence and structure such as GSPHI [65], GCN-based approach [56] and RBP-based approach [54], databases of protein information were required. Commonly used databases for protein include UniprotKB [66], UniRef [67] and MillardLab, which collected the protein function data, protein structure information, protein–protein interaction data and biological pathway information. Some computational methods such as RaFAH and Leite's study also analyze the similarity of protein domain between phages and bacteria, which were collected from the protein family database Pfam [68].

DESCRIPTOR FOR PHI PREDICTION

Computational methods require descriptors as an indication for PHI prediction, which establish the connection between phage and its host. For example, BLAST-based approach compared the similarity of homologous sequences and CRISPR-based approach compared the similarity between CRISPR spacers and viral genomes, both of which were known genome segments and can be found in the genome of phage and its host. However, some of such specific genome sequences between phages and

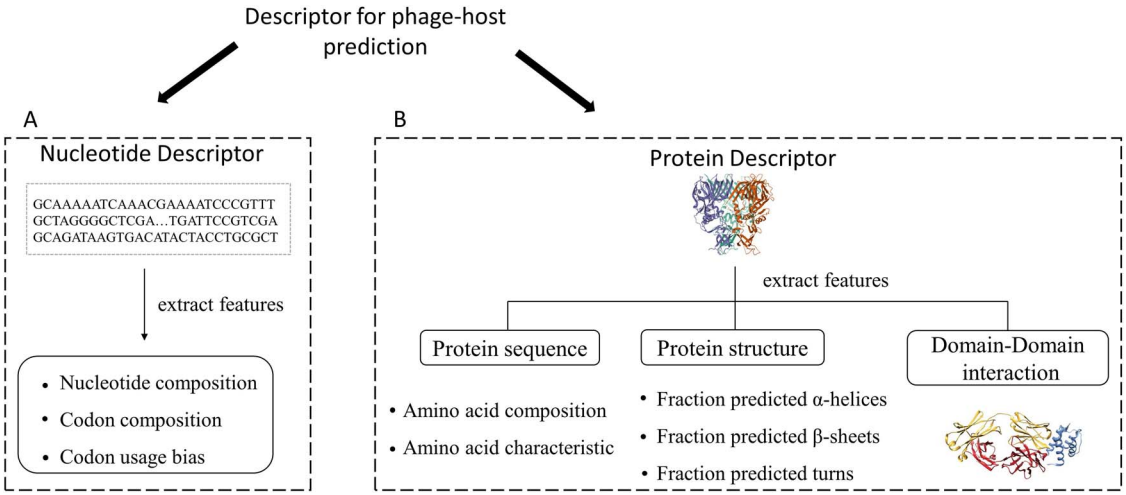

Figure 3. Typical descriptors in the computational method for phage–host prediction. Descriptors applied in PHI prediction can be classified into nucleotide descriptor and protein descriptor. (**A**) Nucleotide descriptor, including nucleotide composition, codon composition and codon usage bias. (**B**) Protein descriptor, including characteristics of amino acid, protein structure and domain–domain interaction.

bacteria have not been detected, without which the prediction of phage–host cannot be achieved, whereas some nucleotide and protein characters such as oligonucleotide frequency, amino acid frequency, protein physicochemical properties are confirmed to yield signals to recognize PHI [69]. Thus, descriptors indicating nucleotide and protein characteristics are also widely employed for predicting PHI, typical descriptors are shown in Figure 3.

Oligonucleotide frequency is considered as the most commonly used nucleotide descriptor for phage–host prediction, which is also known as k-mer frequency. It was first proposed by Woese and Fox [36], and was first applied in PHI identification by Villarroel [40], who built a bioinformatic tool called HostPhinder. Previous study has verified 16-mers as the optimal k-mer to identify bacterial species [70], HostPhinder therefore predicted bacterial host for phages based on the similarity of 16-mers frequency of viral genome. After that, many other tools based on k-mer frequency were invented. For example, VirusHostMatcher compared the dissimilarity of oligonucleotide frequency between bacteria and phages; it constructed a model based on k-mer of different lengths and optimized 6-mers as the best parameter in connecting viral and host genomes. Moreover, Edward illustrated that compared performance of k-mer with long length result in highly specific oligonucleotide, and screened 4-mers to be optimal k-mer by calculating k-mer similarity of length from 3 to 8 bp between viral and bacterial genome to predict PHI. What is more, WIsH, which is currently the best tool for host prediction of phages based on short phage contig, applied 8-mers as feature. Therefore, the optimal length of k-mer differs from phage–host prediction approaches with different machine learning models. Numerous researches indicate that the k-mer length is influential for the prediction performance. Shorter k-mer shows low specificity, on which bacteriophage–host interaction prediction based performed worse, whereas the prediction relies on longer k-mer which is too specific and results that the host of the query bacteria cannot be predicted at all. Previous studies also confirmed that the measurement of k-mer dissimilarity has the potential to influence prediction performance [41]. Although PHI prediction based on k-mer composition shows compellent performance, it requires bacteria and phages with whole-genome sequence as long as possible to obtain the profile that is representative of the genome. Furthermore, other nucleotide characteristics such as GC-content [71] and codon usage bias [72] have been verified as signals of the relationship between phages and their hosts' genomes, which can also be used as descriptors to construct prediction model [54].

Not only do infection mechanisms of phages make protein of receptor-binding and lysins a link between phages and bacteria, but bacterial defense mechanisms establish protein relationships between phages and bacteria [73]. Thus, some methods also take protein characteristic as the descriptor to construct PHI prediction tools [74]. Features such as amino acid abundance, protein secondary structure, physicochemical properties and domain–domain interaction have been verified with indicative activity in PHI. For example, Leite's study identified potential hosts using domain–domain interaction along with protein primary structure features including the frequency of amino acid, chemical elements composing the protein and the molecular weight of the protein. Since the applicability of nucleotide character and protein features in PHI identification, some tools combine protein and nucleotide sequences to achieve comprehensive analysis and have shown satisfactory performance [54, 56].

Except for constructing handcraft features based on genomic and protein sequences, the embeddings generated by protein language models have been demonstrated with better prediction performance for phages and hosts with known RBPs [58].

MODELS APPLIED IN PHI PREDICTION

Machine learning approaches were widely employed in current computational methods of PHI prediction, which can be divided into traditional machine learning methods and deep learning methods. Typical models of current computational methods are listed in Table 3.

For methods based on k-mer similarity, machine learning methods were widely applied for PHI prediction, which can be classified into supervised learning, semi-supervised and unsupervised learning methods. Bioinformatic prediction tools for PHI based on the unsupervised learning are dependent on unlabeled data. For example, PhagehostPredictor [42] applied the Gaussian model along with the k-means Algorithm to achieve the host prediction for query phages based on the frequency similarity of 4-mers, HostPhinder [40] applied clustering algorithm to identify

Table 3: Machine learning methods in phage–host prediction tools

Prediction tool	Descriptor	Machine learning method	Main strength
HostPhinder [48]	Nucleotide 16-mer frequencies	Clustering algorithm	Methods based on traditional machine learning methods are interpretable and easy for users to understand.
VHM [49]	Nucleotide 6-mer frequencies		
WIsH [54]	Nucleotide 8-mer frequencies	Markov model	
PHP [50]	Nucleotide 4-mer frequencies	Gaussian model	
RBP [61]	Properties of RBP	Random Forest	
RaFAH [39]	Custom HMM profiles	Random Forest	
PHERI [59]	Properties of protein sequences	Binary decision tree Classifier	
LMFH-VH [55]	Nucleotide 6-mer frequencies	Heterogeneous network	Approaches based on deep learning methods can show better performance with accumulating genome data of phages and their hosts
ILMF-VH [56]	Nucleotide 6-mer frequencies	Heterogeneous network	
Leite's study [57]	Properties of protein sequences	Artificial Neural Networks	
PredPHI [58]	Properties of protein sequences	Deep Convolutional Neural Network	
GCN [65]	Multiple features	Graph convolutional network	
GSPHI [71]	Properties of RBP	Deep Convolutional Neural Network	
vHULK [81]	Protein families (pVOGs)	Deep Convolutional Neural Network	

the host for query phages. Conversely, some methods are based on supervised learning trained machine to infer the PHI by analyzing features of known phage–host pairs and phage–nonhost pairs, which are more dependent on the labeled data. RaFAH, PHERI applied algorithm of Random Forest, Decision Tree, which are considered as Classification Algorithm in supervised learning. Meanwhile, the Regression Algorithm was also employed along with the Classification Algorithm in Boeckaerts's study to identify the interaction between phages and bacteria. Furthermore, semi-supervised learning is also applied in the host prediction of phage. HostG [56] can conduct host prediction for novel viruses by analyzing virus–virus protein similarity and virus–host DNA sequence similarity with GCN.

For computational methods relying on protein analysis or multiple features including nucleotide sequences and protein sequences, deep learning methods were commonly applied. Extensively accumulated information refers to PHI and genomic sequences of phages and bacteria in the era of big data sets facilitate deep learning methods to predict PHI. PredPHI [51] used deep convolutional neural network to construct the bioinformatic model based on the information of protein–protein interaction between bacteriophage and hosts. Meanwhile, LMFH-VH [46] and ILMF-VH [47] applied network similarity fusion and heterogeneous networks to connect genomic similarity between phage–phage, virus–virus and phage–virus, and achieved the interaction prediction along with kernelized logistic matrix factorization algorithm. Leite's study [50] for interacting phage–bacteria pairs employed the artificial neural network as the optimistic machine-learning modeling technique. However, these network-based approaches ignored the sparsity of phage–host information network, some models were therefore applied to address the problem. For example, PHIAF [57] utilized GAN model to achieve data augmentation, and GERMAN-PHI employed Graph Embedding Representation learning with Multi-head Attention mechanism for PHI prediction, both of which illustrated comparable performance.

Different from the methods that based on handcraft features, a new direction that generating dense vector representations of specific proteins by protein language model is proposed [58]. According to this research, combining handcrafted features with the dense vector representations did not significantly increase prediction performance, which suggested that the embeddings-based model captured features with combination of genomic and DNA characteristics. Therefore, the protein language model may be a potential substitute for constructed descriptors of nucleotides and proteins between phages and host bacteria.

PREDICTION PERFORMANCE AND APPLICATION SCOPE

On overviewing of PHI prediction researches, measurements including recall, F-score, accuracy, ROC-AUC and PR-AUC are widely applied to evaluate performance for phage–host prediction methods, and the claimed prediction performance of computational methods has been listed in Table 4.

It is also concluded that some features such as sequence length, reference database and the selection of negative samples can significantly influence prediction performance. Most tools can only show good prediction performance for the bacteriophage and host with whole-genome sequence, which consists of long contigs, yet phages and bacteria with short contigs widely existed especially in the metagenome. Furthermore, current tools show worse prediction performance on the strict species level, as taxonomic information of hosts are on genus or species level in the current PHIs available in the database. However, treatment of bacterial infection with phage therapy, especially the treatment applied in the clinic, requires to utilize phage with specific host strains instead of bacteria on the same taxonomic level from genus to kingdom with the infected bacteria. In addition, to the best of our knowledge, there was no experimentally verified phage–nonhost pair in any currently available databases. Therefore, these negative samples were randomly selected in some researches of PHI prediction [50], which may lead to unstable prediction performance. Some other methods constructed custom negative samples by defining nonhost bacteria as the one with different taxonomy from host bacteria. However, there are still some host bacteria in the custom negative samples on consideration of cross-reaction cases, which means some bacteria are able to infect bacteria on different genus levels or even species levels [20]. Therefore, optimized means of selecting exact negative samples are required for convincing prediction performance.

Except for the above factors that influence performance of trained models, some other factors related to computational complexity are also concluded (Table 5), and the PHI prediction tools have been classified according to these influential factors (Supplementary Table 1). First, the learning algorithms utilized in PHI tools can be classified into statistical model, traditional machine learning model and deep learning model. Theoretically,

Table 4: Performance of phage–host prediction tools

Phage–host prediction tool	Prediction performance						Data set
	Taxonomy	Accuracy	Recall	F-score	ROC-AUC	PR-AUC	
VHM [41]	Genus	64%	–	–	–	–	Data set of 820 bacteriophages and 2699 bacterial genomes
PHP [42]	Genus	35%	100%	–	–	–	Test data set of 671 viral genomes and 60 105 prokaryotic genomes
RaFAH [31]	Genus	Balance of high precision and recall[a]					Viral RefSeq genomes published after October 2019
PHERI [52]	Genus	–	82%	–	–	–	Test data set of 1202 phage sequences that infect 50 bacterial genera
WIsH [45]	Genus	63%	–	–	87%	–	Benchmark of 1420 viral genomes and 3780 prokaryotic genomes
HostG [56]	Genus	70%	–	–	–	–	Test data set of 671 pairs of viral genomes and 60 105 prokaryotic genomes
HostPhinder [40]	Species	74%	–	–	–	–	Phage set of 1546 phage genomes
RBP [54]	Species	–	–	–	–	74–94%	Data set consisted of 887 RBPs related to seven bacterial host
LMFH-VH [46]	Species	63%	–	–		–	Benchmark of 820 viruses and 2699 host genomes
ILMF-VH [47]	Species	64%	–	–		–	Benchmark of 820 viruses and 2699 host genomes
Leite's study [50]	Species	86–98%	85–98%	86–97%		–	Data sets of 1064 phage–bacterium pairs Data sets of 1064 phage–bacterium pairs
PredPHI [51]	Species	–	–	–	81%	78%	Genome data set of 618 phage–host pairs and 618 phage–nonhost pairs
GSPHI [65]	Species	87%	–	–	92%	–	ESKAPE data set of 1232 PHI pairs consisting of nine bacterial species
vHULK [55]	Species	52%	–	–	–	–	Test data set of 2153 phage genomes
PHIAF [57]	Genus	–	–	–	88%	86%	Genome and proteome set of 312 PHIs between 304 phage and 235 hosts.
Gonzales' study [58]	Genus	–	27–60%	25–63%	–	–	Genome and protein data set of 24 752 RBPs across 9583 phages and 232 hosts.
GERMAN-PHI [48]	Genus	–	–	–	88%	–	Data set consists of 1436 PHIs between 1330 phages and 171 hosts.

[a]Compared with peer review methods, RaFAH shows higher *F*-score, which means balance of high precision and recall. However, it did not show exact value of *F*-score on genus level.

the method-based deep learning model is more complex than that based on statistical model, traditional machine learning model, as the deep learning models have more parameters, and require more data to train. In addition, the deep learning model strongly relies on computing power, which requires specialized hardware (e.g. GPUs). Therefore, the methods based on statistical model and traditional machine learning model performed better on the computational complexity. In addition, the complexity of input data is also a considerable factor to rank the PHI tools. Considering the difficulty level of data collection, the PHI tool that require genomic data of query phages only is the most convenient, as these tools provided reference bacteria database. Although such methods are convenient to use, the reference database needs to be updated and enriched consequently by the developer. The second level is the tools that require data of both phages and bacteria, and available databases that collected numerous viral and bacterial genome and proteome have been summarized in Table 2 of the review manuscript. Except for the above methods, some PHI tools ask the users to input specific proteins of PHI, which cannot be acquired easily from available databases, and the high demand of input data limited the application scope of phage–host prediction. Furthermore, some PHI tools, especially the tools that take proteome data as input, require extra tools such as DIAMOD, Prodigal, Prokka and HMMER to predict and transfer genome sequence into coding sequence and domain features, aggravating the computational complexity.

To better compare the effectivity and computational complexity of different PHI tools, six testable PHI tools were screened and the performance of these tools was testified based on a benchmark set. The set is provided in Li's study [57], and the data exist in the training sets of the six PHI methods was discarded, leading to a final set with 90 phage–host pairs and 90 phage–nonhost pairs (Supplementary Tables 2 and 3), along with genome and proteome data of related phages and bacteria. The result of these PHI tools shows that the tools based on extra tools need more time to predict PHI, as the running time of RaFAH and vHULK for predicting a pair of phage–bacteria is longer than other methods without relying on extra tools (Table 6). In addition, VHM showed the best performance with accuracy of 0.678, followed by the PB-LKS and WIsH method, indicating that these methods have better ability to correctly predict both phage–host pairs and phage–nonhost pairs compared with other methods. Among these three methods, PB-LKS shows balanced values on each measurement, whereas the accuracy of VHM and WIsH is led by the high specificity but lower sensitivity. In other words, for users require to differ PHIs and phage–nonhost interactions, the PB-LKS is better than other methods. Meanwhile, the other five methods show

Table 5: The influence factors of PHI tools refer to computational complexity

PHI tools	Learning model	Input data	Dependent tools
VHM [41]	–	Phage and bacterial genome	–
HostPhinder [40]	–	Phage genome	–
WIsH [45]	Statistical model	Phage and bacterial genome	–
PHP [42]	Statistical model	Phage genome	–
PB-LKS [49]	Simple machine learning model	Phage and bacterial genome	–
RBP [54]	Simple machine learning model	RBP	–
RaFAH [31]	Simple machine learning model	Phage genome	Prodigal, HMMER
PHERI [52]	Simple machine learning model	Phage genome	Prodigal, Prokka
LMFH-VH [46]	Deep learning model	Phage and bacterial genome	–
ILMF-VH [47]	Deep learning model	Phage and bacterial genome	–
PredPHI [51]	Deep learning model	Phage and bacterial proteome	–
GERMAN-PHI [48]	Deep learning model	Phage and bacterial genome	–
GSPHI [65]	Deep learning model	RBP	–
HostG [56]	Deep learning model	Phage genome	DIAMOND, Prodigal
Leite's study [50]	Deep learning model	Phage and bacterial proteome	GeneMarks, HMMER
vHULK [55]	Deep learning model	Phage genome	Prokka, HMMER
PHIAF [57]	Comprehensive model	Phage and bacterial genome and proteome	–
Gonzales's study [58]	Comprehensive model	RBP	–

Table 6: The performance of PHI tools based on the benchmark data set in defaulted threshold

PHI tools	Feature	Runtime (Per P-B pair)	Accuracy	Sensitivity	Specificity	Precision	F-score
PB-LKS [49]	K-mer composition	≈ 10 s	0.656	0.622	0.689	0.667	0.644
PHP [42]	K-mer composition	≈ 10 s	0.539	0.078	1.000	1.000	0.144
VHM [41]	K-mer composition	≈ 10 s	0.678	0.389	0.967	0.921	0.547
WIsH [45]	K-mer composition	≈ 10 s	0.606	0.244	0.967	0.880	0.383
RaFAH [31]	Protein	≈ 1 min	0.483	0.122	0.844	0.440	0.191
vHULK [55]	Protein	≈ 10 min	0.428	0.122	0.733	0.314	0.176

better specificity with worse sensitivity, which illustrates that most of the predicted phage–host pairs are correct, whereas few phage–host pairs can be recognized from the 90 phage–host pairs. And the PHP tool shows best specificity and precision, which suggests that it shows great performance on correctly predicting PHI. And for phage therapy that requires to correctly select infectious phage for pathogenic bacteria, such tools with high prediction accuracy are recommended. Furthermore, the PHI tools of RaFAH and vHULK performed worse on both prediction performance and running time base on the constructed benchmark set. The consuming time is led by the extra relied tools to transfer the genome sequence into protein features. And the possible reason of unsatisfied performance is that the nucleotide features are more important than protein features in identifying PHI, which is also proposed in previous study [57].

As the improvement of performance, computational method has been widely applied in the characterization of the newly discovered virus [75]. Meanwhile, these methods are also used to characterize viruses from metagenomes such as forest soil metagenome [76] and gut metagenome [77]. Predicting PHI also shows potential in phage therapy, which has been utilized in phage therapy against bacterial infection [78]. As a promising candidate to combat antimicrobial resistance, phage therapy has been known for its potential clinical utility in lysing bacteria [79]. The narrow specificity of phage allows it to kill pathogenic bacteria without influencing commensal organisms and healthy microbiomes. As a result, phage therapy has been applied as treatment of urinary tract infections [80], skin and soft tissue infections [81] and gastrointestinal infections [82] related to pathogens such as *Staphylococcus aureus*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*. Consequently, applying high-throughput computational methods for PHI with high prediction accuracy can promote the development of phage therapy against bacterial infection.

DISCUSSION AND CONCLUSION

Identifying hosts of novel phages has been confirmed with applications of gene transfer search [83], disease diagnosis [84] and bacteria detection [85]. Experimental method is still the most convincing approach for PHI identification. However, <1% of microbial hosts have been successfully cultivated with laboratory condition [86]. Thus, applying computational methods with good prediction accuracy to identify PHI is currently an alternative to experimental method, showing great potential for targeted phage therapy. With the rapid development of sequencing technology, the accumulation of phages and bacterial genome sequences has provided great convenience for computational methods. All of the computational methods show their specific advantages and disadvantages, combining different computational methods is possible to improve the performance of existing computational prediction methods. On overviewing of PHI prediction researches, it is confirmed that alignment-based methods always show better prediction accuracy but narrow applicability than alignment-free methods. The probable reason is that homologous sequences that arose with the coevolution of phages and host bacteria can

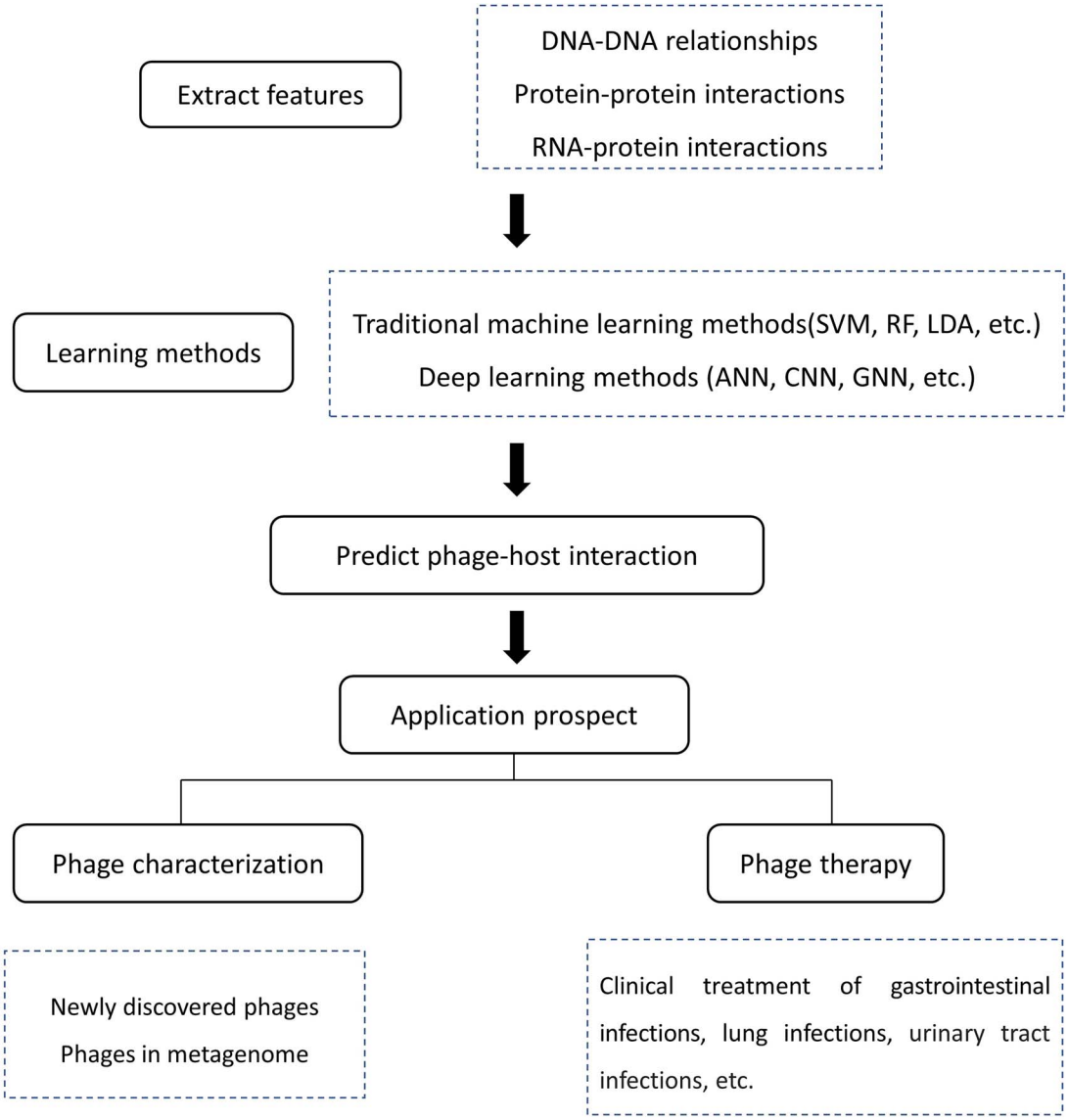

Figure 4. Prospective research of computational methods for PHI prediction. Descriptor features including DNA, RNA and protein and mutual interactions can be further involved in future research, and the optimized algorithm will be selected from traditional machine learning methods and deep learning methods to improve prediction performance. Future application is not only limited to phage characterization, but also phage therapies against bacterial infection.

be accurately detected with alignment-based method, whereas these homologous sequences exist in a few phages and bacteria. In addition, the approaches base on neural network can show better performance with accumulating genome data of phages and their hosts, but the black-box of these methods makes it hard to interpret.

Nevertheless, current computational methods have shown some common limitations. First, the dependence of these prediction methods on the genome sequences which are limited in the reference database restricts the prediction accuracy. Although numerous databases have collected experimentally testified PHI currently, phages and bacteria related to some genera are still scarce. From the host distribution of a common-used training data set in the phage–host prediction methods which includes 1426 pairs of PHI compiled from the NCBI RefSeq database before 5 May 2015, it is found that there are 43 host genera that consist of only one host bacteria strain [49], which will result in the inaccurate prediction performance of phage–host related to these host genera.

Furthermore, selected nonhost bacteria utilized in current computational methods have not been experimentally testified, and there may be potential phage–host pairs among all these phage–nonhost pairs; therefore, positive and unlabeled learning algorithm [87] is potential for constructing PHI model based on such training set consists of experimentally verified positive samples and unverified negative samples. Currently, some methods have defined the nonhost bacteria as the bacteria at different genus level with the host bacteria to overcome the negative sample limitation, but minimal host bacteria may not be removed from data set of nonhost bacteria because of few multi-hosts phages, leading to unfaithful prediction performance. Thus, the experimental method for phage–host identification is still a golden standard. Nevertheless, for phage–host prediction tools employed in the phage therapy against bacterial infection,

the missing potential 'true positive' is not as important as ensuring the predicted 'positive samples' are correct [49]. In this regard, the limitation of experimentally testified nonhost bacteria may not significantly influence the prediction performance. As the potential application in the phage therapy against bacterial infection, another limitation for PHI is that these methods cannot be performed on the strain level, whereas the phage therapy requires to discern mutant bacteria from candidate host bacteria strains.

As a promising approach for bacteriophage–host interaction identification, computational methods show great potential for pre-screening candidate hosts in further experimental verification and application of phage therapy, the perspective of future research on PHI is described in Figure 4. Current researches concentrate mainly on nucleotide relationship and protein–protein interaction between phage and bacteria to connect host bacteria with virus. Nevertheless, a single-standard RNA SARS-CoV-2 has been reported with possible interaction with host protein [88]. Therefore, combining RNA-protein interaction along with nucleotide and protein characteristics may be supported to improve prediction performance for future computation methods. What is more, currently constructed bioinformatic tools can predict comprehensive PHI, whereas clinical bacterial infection mainly involves virulent bacteriophages. Clarifying the infection mechanism of virulent bacteriophages and developing a computational model to specifically predict host for virulent phages can foster the application of these predictive *in silico* methods in phage therapy. In future research, these computational methods can be combined with artificial intelligence model and pretrained model for natural language to achieve better promotion of performance. The continuous improvement of computational methods will contribute to the understanding of PHIs in natural systems and the application of phage therapy against bacterial infection. For the application in phage therapy that requires to discern mutants from candidate host bacteria strains, RBP-based methods [54, 58] show great potential. The PHI can be further developed by targeting specific RBPs of exact phage–host infection mechanism, and correspondingly construct prediction model with strong selectivity and high accuracy to facilitate the development of computational phage therapy. Note that, the current computational methods rely on different known mechanisms between PHI, such as the lysogenic bacteriophage infection [89], CRISPR system [17] and specific gene integration [18]. In the future, if new mechanisms were detected between PHIs, it can be used to guide the new generation of *in silico* models for multiple scenarios.

Key Points

- Predictive accuracy of alignment-based methods and alignment-free methods for phage–host interaction (PHI) prediction validates convinced correlation between phages and their hosts, which come from homologous sequences integrated into bacterial genome and coevolution of evolutionary arms-race between phages and their hosts.
- Commonly used reference databases and related feature descriptors which contribute to the construction of computational approaches in PHI prediction are concluded. Future directions of computational researches can be inspired by the summarization.
- Phage therapy has been confirmed with applicability in the treatment of numerous bacterial infections with the basic that phages can specifically recognize bacteria host and lyse bacteria. High-throughput *in silico* model with good performance to predict PHI shows prospective application in future phage therapy.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENT

Supported by the Medical Science Data Center of Fudan University.

AUTHORS' CONTRIBUTIONS

W.N. and T.Q. contributed to the conception of the study and wrote the manuscript. Y.W. and H.D. tested the current available PHI tools based on the same benchmark data set. Z.G. modified the manuscript. J.Q. supervised the whole project.

FUNDING

This work was supported by grants from the National Key Research and Development Program of China (2022YFF1101100), the National Natural Science Foundation of China (32000470), the National Natural Science Foundation of China (32370697).

DATA AVAILABILITY

For access to any research-related data, kindly reach out to the corresponding author.

REFERENCES

1. Twort FW. An investigation on the nature of ultra-microscopic viruses. *Lancet* 1915;**186**(4814):1241–3.
2. D'Herelle MF. Sur un microbe invisible antagoniste des bacilles dysenteriques. *CR Acad Sci Ser D*. 1917;**165**:373–5.
3. Schofield DA, Sharp NJ, Westwater C. Phage-based platforms for the clinical detection of human bacterial pathogens. *Bacteriophage* 2012;**2**(2):105–21.
4. Bao Q, Li X, Han G, *et al*. Phage-based vaccines. *Adv Drug Deliv Rev* 2019;**145**:40–56.
5. Fenton M, Ross P, McAuliffe O, *et al*. Recombinant bacteriophage lysins as antibacterials. *Bioeng Bugs* 2010;**1**(1):9–16.
6. Jassim SA, Limoges RG, El-Cheikh H. Bacteriophage biocontrol in wastewater treatment. *World J Microbiol Biotechnol* 2016;**32**(4):70.
7. Kortright KE, Chan BK, Koff JL, Turner PE. Phage therapy: a renewed approach to combat antibiotic-resistant bacteria. *Cell Host Microbe* 2019;**25**(2):219–32.
8. Edwards RA, McNair K, Faust K, *et al*. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 2016;**40**(2):258–72.
9. Lenski RE. Dynamics of Interactions between Bacteria and Virulent Bacteriophage. In: Marshall KC (ed). *Advances in Microbial Ecology*. Boston, MA: Springer US, 1988, 1–44.
10. Rostol JT, Marraffini L. (Ph)ighting Phages: how bacteria resist their parasites. *Cell Host Microbe* 2019;**25**(2):184–94.

11. Folimonova SY. Superinfection exclusion is an active virus-controlled function that requires a specific viral protein. *J Virol* 2012;**86**(10):5554–61.
12. Ofir G, Melamed S, Sberro H, *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat Microbiol* 2018;**3**(1):90–8.
13. Hille F, Richter H, Wong SP, *et al.* The biology of CRISPR-Cas: backward and forward. *Cell* 2018;**172**(6):1239–59.
14. Tal N, Sorek R. SnapShot: bacterial immunity. *Cell* 2022;**185**(3):578–578.e1.
15. Gao Z, Feng Y. Bacteriophage strategies for overcoming host antiviral immunity. *Front Microbiol* 2023;**14**:1211793.
16. Stern A, Sorek R. The phage-host arms race: shaping the evolution of microbes. *Bioessays* 2011;**33**(1):43–51.
17. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 2010;**327**(5962):167–70.
18. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* 2002;**30**(4):866–75.
19. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
20. Pedersen JS, Carstens AB, Djurhuus AM, *et al.* Pectobacterium Phage Jarilo displays broad host range and represents a novel genus of bacteriophages within the family Autographiviridae. *Phage (New Rochelle)* 2020;**1**(4):237–44.
21. Barrangou R, Fremaux C, Deveau H, *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;**315**(5819):1709–12.
22. Marraffini LA. CRISPR-Cas immunity in prokaryotes. *Nature* 2015;**526**(7571):55–61.
23. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. *Bioinformatics* 2005;**21**(Suppl 1):i152–8.
24. Biswas A, Staals RHJ, Morales SE, *et al.* CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* 2016;**17**:356.
25. Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**(7402):207–14.
26. Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol Ecol* 2011;**77**(1):120–33.
27. Sanguino L, Franqueville L, Vogel TM, Larose C. Linking environmental prokaryotic viruses and their host through CRISPRs. *FEMS Microbiol Ecol* 2015;**91**(5):fiv046.
28. Dion MB, Plante PL, Zufferey E, *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* 2021;**49**(6):3127–38.
29. Zhang R, Mirdita M, Levy Karin E, *et al.* SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* 2021;**37**(19):3364–6.
30. Makarova KS, Wolf YI, Iranzo J, *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 2020;**18**(2):67–83.
31. Coutinho FH, Zaragoza-Solas A, López-Pérez M, *et al.* RaFAH: host prediction for viruses of bacteria and archaea based on protein content. *Patterns (N Y)* 2021;**2**(7):100274.
32. Stern A, Mick E, Tirosh I, *et al.* CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* 2012;**22**(10):1985–94.
33. Hingamp P, Grimsley N, Acinas SG, *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* 2013;**7**(9):1678–95.
34. Silva GG, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2014;**2**:e425.
35. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000;**7**(1–2):203–14.
36. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* 1977;**74**(11):5088–90.
37. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;**27**(6):764–70.
38. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**(3):R46.
39. Edwards RA, Olson R, Disz T, *et al.* Real time metagenomics: using k-mers to annotate metagenomes. *Bioinformatics* 2012;**28**(24):3316–7.
40. Villarroel J, Kleinheinz K, Jurtz V, *et al.* HostPhinder: a phage host prediction tool. *Viruses* 2016;**8**(5):116.
41. Ahlgren NA, Ren J, Lu YY, *et al.* Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 2017;**45**(1):39–53.
42. Lu C, Zhang Z, Cai Z, *et al.* Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021;**19**(1):5.
43. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015;**3**:e985.
44. Smits SL, Bodewes R, Ruiz-Gonzalez A, *et al.* Assembly of viral genomes from metagenomes. *Front Microbiol* 2014;**5**:714.
45. Galiez C, Siebert M, Enault F, *et al.* WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 2017;**33**(19):3113–4.
46. Liu D, Hu X, He T, Jiang X. Virus-host association prediction by using kernelized logistic matrix factorization on heterogeneous networks. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Madrid, Spain, 2018, pp. 108–13.
47. Liu D, Ma Y, Jiang X, He T. Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* 2019;**20**(Suppl 16):594.
48. Wang Y, Sun H, Wang H, *et al.* An effective model for predicting phage-host interactions via graph embedding representation learning with multi-head attention mechanism. *IEEE J Biomed Health Inform* 2023;**27**(6):3061–71.
49. Qiu J, Nie W, Ding H, *et al.* PB-LKS: a python package for predicting phage-bacteria interaction through local k-mer strategy. *Brief Bioinform* 2024;**25**(2):bbae010.
50. Leite DMC, Brochet X, Resch G, *et al.* Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 2018;**19**(Suppl 14):420.
51. Li M, Wang Y, Li F, *et al.* A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**(5):1801–10.
52. Baláz A, *et al.* PHERI-Phage Host ExploRation Pipeline. *Microorganisms*. 2023;**11**(6):1398.
53. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**(7):1575–84.
54. Boeckaerts D, Stock M, Criel B, *et al.* Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci Rep* 2021;**11**(1):1467.
55. Amgarten D, Iha BKV, Piroupo CM, *et al.* vHULK, a new tool for bacteriophage host prediction based on annotated genomic

features and neural networks. *Phage (New Rochelle)* 2022;**3**(4):204–12.
56. Shang J, Sun Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol* 2021;**19**(1):250.
57. Li M, Zhang W. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Brief Bioinform* 2022;**23**(1):bbab348.
58. Gonzales MEM, Ureta JC, Shrestha AMS. Protein embeddings improve phage-host interaction prediction. *PloS One* 2023;**18**(7):e0289030.
59. Sayers EW, Agarwala R, Bolton EE, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2019;**47**(D1):D23–8.
60. Gao NL, Zhang C, Zhang Z, *et al.* MVP: a microbe-phage interaction database. *Nucleic Acids Res* 2018;**46**(D1):D700–7.
61. Lamy-Besnier Q, Brancotte B, Ménager H, Debarbieux L. Viral host range database, an online tool for recording, analyzing and disseminating virus-host interactions. *Bioinformatics* 2021;**37**(17):2798–801.
62. Mihara T, Nishimura Y, Shimizu Y, *et al.* Linking virus genomes with host taxonomy. *Viruses* 2016;**8**(3):66.
63. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics* 2017;**33**(5):784–6.
64. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
65. Pan J, You W, Lu X, *et al.* GSPHI: a novel deep learning model for predicting phage-host interactions via multiple biological information. *Comput Struct Biotechnol J* 2023;**21**:3404–13.
66. Boutet E, Lieberherr D, Tognolli M, *et al.* UniProtKB/Swiss-Prot. *Methods Mol Biol* 2007;**406**:89–112.
67. Suzek BE, Huang H, McGarvey P, *et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;**23**(10):1282–8.
68. Mistry J, Chuguransky S, Williams L, *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**(D1):D412–9.
69. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. *PLoS Comput Biol* 2020;**16**(5):e1007894.
70. Larsen MV, Cosentino S, Lukjancenko O, *et al.* Benchmarking of methods for genomic taxonomy. *J Clin Microbiol* 2014;**52**(5):1529–39.
71. Swan BK, Tupper B, Sczyrba A, *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* 2013;**110**(28):11463–8.
72. Carbone A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* 2008;**66**(3):210–23.
73. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 2010;**8**(5):317–27.
74. Coelho ED, Arrais JP, Matos S, *et al.* Computational prediction of the human-microbial oral interactome. *BMC Syst Biol* 2014;**8**:24.
75. Lopez MES, Gontijo MTP, Cardoso RR, *et al.* Complete genome analysis of Tequatrovirus ufvareg1, a Tequatrovirus species inhibiting Escherichia coli O157:H7. *Front Cell Infect Microbiol* 2023;**13**:1178248.
76. Wu R, Smith CA, Buchko GW, *et al.* Structural characterization of a soil viral auxiliary metabolic gene product - a functional chitosanase. *Nat Commun* 2022;**13**(1):5485.
77. Leleiwi I, Rodriguez-Ramos J, Shaffer M, *et al.* Exposing new taxonomic variation with inflammation - a murine model-specific genome database for gut microbiome researchers. *Microbiome* 2023;**11**(1):114.
78. Aggarwal S, Dhall A, Patiyal S, *et al.* An ensemble method for prediction of phage-based therapy against bacterial infections. *Front Microbiol* 2023;**14**:1148579.
79. Hitchcock NM, Devequi Gomes Nunes D, Shiach J, *et al.* Current clinical landscape and global potential of bacteriophage therapy. *Viruses* 2023;**15**(4):1020.
80. Leitner L, Ujmajuridze A, Chanishvili N, *et al.* Intravesical bacteriophages for treating urinary tract infections in patients undergoing transurethral resection of the prostate: a randomised, placebo-controlled, double-blind clinical trial. *Lancet Infect Dis* 2021;**21**(3):427–36.
81. Jault P, Leclerc T, Jennes S, *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**(1):35–45.
82. Febvre HP, Rao S, Gindin M, *et al.* PHAGE study: effects of supplemental bacteriophage intake on inflammation and gut microbiota in healthy adults. *Nutrients* 2019;**11**(3):666.
83. Fernandez L, Rodriguez A, Garcia P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J* 2018;**12**(5):1171–9.
84. Bazan J, Calkosinski I, Gamian A. Phage display—a powerful technique for immunotherapy. *Hum Vaccin Immunother* 2012;**8**(12):1829–35.
85. Edgar R, McKinstry M, Hwang J, *et al.* High-sensitivity bacterial detection using biotin-tagged phage and quantum-dot nanocomplexes. *Proc Natl Acad Sci U S A* 2006;**103**(13):4841–5.
86. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005;**3**(6):504–10.
87. Denis F, Gilleron R, Letouzey F. Learning from positive and unlabeled examples. *Theor Comp Sci* 2005;**348**(1):70–83.
88. Kim D, Lee JY, Yang JS, *et al.* The architecture of SARS-CoV-2 transcriptome. *Cell* 2020;**181**(4):914–921.e10.
89. Davies EV, Winstanley C, Fothergill JL, James CE. The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol Lett* 2016;**363**(5):fnw015.