


PB-LKS: a python package for predicting phage–bacteria interaction through local K-mer strategy

Jingxuan Qiu[†], Wanchun Nie[†], Hao Ding, Jia Dai, Yiwen Wei, Dezhi Li, Yuxi Zhang, Junting Xie, Xinxin Tian, Nannan Wu and

Tianyi Qiu 

Corresponding authors: Tianyi Qiu, Institute of Clinical Science, Zhongshan Hospital, Shanghai Institute of Infectious Disease and Biosecurity, Intelligent Medicine Institute, Fudan University, No.180, Fenglin Road, Xuhui District, Shanghai 200032, China. Tel./Fax: 021-64041990; E-mail: ty_qiu@126.com; Nannan Wu, Shanghai Institute of Phage, Shanghai Public Health Clinical Center, Fudan University, No.2901, Caolang Highway, Jinshan District, Shanghai 201508, China. Tel./Fax: 021-37990333; E-mail: wunannan@shphc.org.cn

[†]Jingxuan Qiu and Wanchun Nie have contributed equally to this work.

Abstract

Bacteriophages can help the treatment of bacterial infections yet require *in-silico* models to deal with the great genetic diversity between phages and bacteria. Despite the tolerable prediction performance, the application scope of current approaches is limited to the prediction at the species level, which cannot accurately predict the relationship of phages across strain mutants. This has hindered the development of phage therapeutics based on the prediction of phage–bacteria relationships. In this paper, we present, PB-LKS, to predict the phage–bacteria interaction based on local K-mer strategy with higher performance and wider applicability. The utility of PB-LKS is rigorously validated through (i) large-scale historical screening, (ii) case study at the class level and (iii) *in vitro* simulation of bacterial antiphage resistance at the strain mutant level. The PB-LKS approach could outperform the current state-of-the-art methods and illustrate potential clinical utility in pre-optimized phage therapy design.

Keywords: phage–bacteria interaction; local K-mer strategy; bioinformatics; machine learning; genome sequence analysis

INTRODUCTION

Bacteriophages (Phages), known as the viruses for bacteria, are the most abundant organism in the biosphere and can be found in all places where their bacterial hosts exist [1]. Current studies showed that the use of phages to cure bacterial infection could be a promising alternative to chemical antibiotics [2]. More importantly, with the rapid emergence of drug-resistant bacteria caused by the overconsumption of antimicrobials, phage therapy shows therapeutic effects for drug-resistant bacteria [3]. The

effectiveness of phage therapy is based on the mechanism that therapeutic phages could specifically lyse the bacteria which caused the disease without harming other commensal bacteria. This is mainly because a phage only infects target bacteria which express its receptor [4]. Lytic phages lyse the host cell to release progeny viruses, which can be a great candidate for the treatment of bacterial infection [5]. Lysogenic phages integrate their nucleic acid into the host cell's DNA or plasmid and replicate without destroying the cell [4]. In addition, a previous study indicated that

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.

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.

Hao Ding is currently a master student at Fudan University. His research interests are bioinformatics and computer science.

Jia Dai is a technician at the Shanghai Institute of Phage and Drug Resistance at the Shanghai Public Health Clinical Center. Her research direction is pathogen biology.

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.

Dezhi Li is a Postdoctoral Researcher at School of Health Science and Engineering, University of Shanghai for Science and Technology. His research interest is phage biology.

Yuxi Zhang 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 machine learning.

Junting Xie is currently a master student at School of Health Science and Engineering, University of Shanghai for Science and Technology. His research interests are bioinformatics and machine learning.

Xinxin Tian is a master student at School of Health Science and Engineering, University of Shanghai for Science and Technology. Her research interests are bioinformatics and antigenicity prediction.

Nannan Wu is an Associate Researcher at the Shanghai Institute of Phage and Drug Resistance at the Shanghai Public Health Clinical Center. His research direction is pathogenic microbiology.

Tianyi Qiu is a professor at the Institute of Clinical Science, Zhongshan Hospital, Fudan University. His research interests are bioinformatics and immunoinformatics.

Received: September 29, 2023. **Revised:** December 16, 2023. **Accepted:** January 5, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

the combined therapy of phage and antibiotics could improve the susceptibility of drug-resistant bacteria to antibiotics and reduce the emergence of resistant bacteria [6]. Meanwhile, the antibacterial spectrum of phages is considerably narrow, which means that a phage tends to have an antibacterial effect on only several or a certain class of bacteria while becoming inefficient for other bacteria. Therefore, the effectiveness of phage therapy relies on the correct match between phage and its host bacteria.

Commonly used experimental approaches to identify phage–bacterial interactions included plaque assays [7], liquid assays [8], viral tagging [9], single-cell sequencing [10] and so on, which were time-consuming and expensive [11] with limited scopes of application to the dramatically increasing number of both phages and bacteria. Therefore, computational approaches for phage–bacteria interaction prediction, which feature high-throughput and culture-independent characteristics, are highly desired [12]. Currently, the computational methods based on metagenomic sequencing and assembly of viral contigs can be roughly divided into alignment-based methods and alignment-free methods [13]. The theoretical basis of alignment-based methods is that the host bacteria usually contain the genomic fragment of the phage [14, 15]. This kind of method relies on the homology sequence alignment between the query phage and host genome, which illustrates high accuracy in the prediction of the phage–bacteria relationship [16]. Among them, BLAST- and CRISPR-based approaches were widely applied. CRISPR-based approaches could obtain higher accuracy than BLAST-based methods but can only be applied to 40–70% of the prokaryotes that encode a CRISPR system [12]. Thus, alignment-based methods are not suitable for the prediction of many novel phage–bacteria relationships.

The alignment-free methods of co-variation analysis link the abundance of co-variation between phage and bacteria sequences across metagenomes in an environment without constructing an explicit homology sequence alignment. This kind of approach can be applied to newly identified phages or bacteria; however, previous studies indicated that the accuracy is relatively lower than approaches involving the sequence homology information [12]. Another alignment-free method is the sequence composition method, which is based on the phenomenon that phages and their hosts often share similar patterns in codon usage or short nucleotide words (K-mer) [17, 18]. Representative works, such as HostPhinder [19], VirHostMatcher (VHM) [20], LMFH-VH [21], ILMF-VH [22], Prokaryotic virus Host Predictor (PHP) [23] and WIsH [24], generated sequence composition features to predict the phage–bacteria interaction. For example, HostPhinder calculated the K-mer similarity of a query phage to each phage with the known host in the reference phage database; the host of the query phage was considered as the host of that phage with most similar K-mer in the database [19]. This means that the prediction accuracy of HostPhinder might be reduced if no phage with high K-mer similarity in the phage database can be found. VHM introduced d_2^* , which is a measurement with background normalization to calculate the oligonucleotide frequency dissimilarity between phages and hosts, while the bacterium with the lowest score is considered the predicted host [20]. With the integrative information from phage–phage, host–host and phage–host association networks, LMFH-VH and ILMF-VH generated the kernelized logistic matrix factorization algorithm based on network similarity fusion and heterogeneous networks for phage–host interaction prediction, respectively. Also, PHP introduced a Gaussian model to calculate the differences in K-mer frequencies between phage and host genome sequence, and the host with the top-ranked score can be considered as a potential host for phage.

The prediction accuracy of the above models at the genus level is ranged from 33% to 58.9% [23]. Furthermore, WIsH [24] is regarded as a useful tool for phages with short contigs, which trained a homogeneous Markov model for each potential host genome and calculates the likelihood of contigs under each model. The one whose model yielded the highest likelihood is considered the host for query phage. For WIsH, the prediction accuracy can reach 63% on the database of 3 kbp phage contigs. However, the prediction accuracy of current alignment-free methods still has room for improvement. This is because the current alignment-free methods rely on the sequence composition features of the whole genome for prediction, which ignores the fact that phage often integrates its genome in the local segments rather than the whole genome sequence of the bacteria genomes. Moreover, the above methods often utilize features from reference databases or interaction networks, in which the prediction fineness can only be achieved at the genus or species level, not for mutants.

The whole genome size of known phages ranged from 2435 bp to 540 kbp, which is smaller than the whole genome size from 112 kbp to 14 Mbp for bacteria [25, 26]. The magnitude differences mean that the gene segment of the phage cannot be evenly integrated into the entire genome of the bacteria, which will lead to the incorrect prediction of alignment-free methods based on the whole genome sequence. Thus, we proposed a local K-mer strategy (LKS) instead of global genome sequence analysis to construct the *in-silico* prediction model for phage–bacteria interaction (PB-LKS). First, the most similar segments (MSSs) in both the phage and bacteria genomes will be screened through the sliding windows before calculating the K-mer frequency of the sequence composition. Then, the K-mer frequencies were used as the descriptors to generate the prediction model with the appropriate algorithm.

The ability of PB-LKS to predict the phage–bacteria interaction is rigorously evaluated on both intra- and interdatasets at different taxonomy levels. It is initially tested through the historical experimentally validated data of 1342 pairs of phage–bacteria relationships, in which PB-LKS could provide a good performance from the kingdom level to the genus level. Then, its ability to pick up positive Antinobacteriophage for bacterial strains from Actinomycetes at the class level was tested on 3455 phage–bacteria interaction pairs derived from PhagesDB [27]. Furthermore, we challenged the utility of PB-LKS at the strain level, to predict the experimentally tested relationships between a previously described phage ϕ Ab124 [28] and nine *Acinetobacter baumannii*. In general, the PB-LKS model could not only provide a high prediction ability to identify the interaction between phages and bacteria strains but also hold the potential to help design clinical phage therapy.

METHODS

LKS-based descriptor generation

This model aimed to construct an alignment-free model to predict phage–bacteria interaction, based on the similarity of co-occurred K-mers. Detailed steps include the followings.

Step1. Local segments splitting. Both the whole genome sequences of phages and bacteria were split into a series of segments using a sliding window. The whole genome sequences consist of a complete genome sequence or several contigs. For the phages and bacteria with a complete genome sequence, the segments were split based on the whole genome sequence, and the whole genome sequence was defined as the split-sequence if the length was shorter than the window size. For genomes that

consist of several contigs which is longer than the window size, the segments were split on each contig, while the segments were defined as the longest contig if the length of every contig is shorter than the window size. The window size ranged from 1000 bp to 15 000 bp. For each window size, three step lengths were tested.

Step2. K-mer frequency calculation for each segment. For each sequence segment of phages and bacteria, the K-mer frequency f_{k-mer} was calculated (K-mer length from 3 to 5). The host vector of f_{k-mer}^{host} and phage vector of f_{k-mer}^{phage} both contained 64-element, 256-element and 1024-element score of all possible 3-nucleotide, 4-nucleotides and 5-nucleotide combination, respectively.

Step3. Screening the MSS. The MSSs between phages and bacteria were defined as those with the largest correlation coefficient of f_{k-mer} . The correlation coefficient was calculated using the *corrcoef* function of numpy 1.22.2 package in Python 3.10.2.

Step4. Calculation of local K-mer descriptor. For the selected MSS, the local K-mer descriptor was defined as the difference between f_{k-mer}^{phage} and f_{k-mer}^{host} .

In the end, the local K-mer descriptor contained the 64-element, 256-element and 1024-element vectors, describing the K-mer difference of the most similar local segment among the whole genome of phage and host.

Algorithm design of PB-LKS

The local K-mer strategy for phage–bacteria interaction prediction involved four steps (Figure 1): (i) divide the genome sequences from both phage and bacteria into segments through sliding windows with defaulted window length (WL) and step size (SS, Figure 1A). (ii) Count the frequency of each K-mer to generate the K-mer frequency descriptors for each segment (Figure 1B). (iii) Rank the correlation coefficient of K-mer descriptors for all of the pairwise segments from the phage and the bacteria to detect the MSS pairs (Figure 1C). (iv) Calculate the local K-mer descriptor defined by the difference between K-mer features from the phage genome and K-mer features from bacteria (Figure 1D). (v) Constructing the PB-LKS model by combining descriptors with appropriate algorithms. The model is designed to provide the interaction prediction for any phage and bacteria with a complete genome sequence (Figure 1E). The dataset used for Model construction and evaluation can be found in the ‘Supplementary Methods Datasets’ part (see Supplementary Data available online at <http://bib.oxfordjournals.org/>), and detailed statistics of the training and test dataset are listed in Supplementary Tables 1 and 2 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

Parameter optimization and model selection of PB-LKS

The key elements of the PB-LKS approach included two essential parts: (i) optimize WL and SS for each segment to generate descriptors for phage and bacteria and (ii) appropriate learning approaches for model construction. Here, we set an initial WL and an SS of 1000 bp, respectively. Then, to obtain the optimized PB-LKS parameters for phage–bacteria interaction prediction, we performed a screening test that traversed the WL ranged from 1000 bp to 15 000 bp with SS set as 0.2 WL, 0.4 WL and 0.8 WL, respectively (Supplementary Table 3, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Furthermore, considering it is a standard binary classification problem that deals with vector information, machine-learning approaches including Bayesnet, Hoeffding Tree, Logistic Regression, Random Tree, Random Forest, XGBoost and Support Vector Machine (SVM), as well as deep-learning approaches of Multi-Layer Perceptron

(MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Bidirectional RNN (Bi-RNN), were introduced for model selection.

RESULTS

Model construction of PB-LKS

The performance of machine learning approaches through 10-fold cross-validation on the initial parameters at the genus level is illustrated in Table 1. Results showed that the model of Random Forest could achieve the highest performance with an ROC-AUC of 0.803 and an accuracy of 0.727, followed by the performance of Logistic regression with an ROC-AUC of 0.780 and an accuracy of 0.712. Considering the better performance of Random Forest compared with all others, it was selected for parameter optimization (see Methods).

Furthermore, a 10-fold cross-validation of the training dataset was provided to evaluate the performance of different parameter combinations through a Random Forest classifier (Supplementary Tables 4–8, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Results showed that increasing the WL at the beginning (1000–9000 bp) can significantly increase prediction performance from 0.770 to 0.860. Meanwhile, with WL greater than 9000 bp, the value of ROC-AUC was maintained to be stable from 0.860 to 0.868, with deviations less than 0.01 (Supplementary Table 4, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The further increase of WL may sometimes decrease the accuracy from 0.775 (WL of 9000) to 0.764 (WL of 10 000) (Supplementary Table 5, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), which means that the WL ranged above 9000 will no further increase the prediction performance. On the other hand, it seems that the step at the first level (0.2 WL) could provide the best prediction performance than the higher SS. Thus, we use the WL of 9000 bp and SS of 1800 bp for optimized parameters to construct the local K-mer descriptor.

The length of K-mers is also an important parameter for phage–host interaction prediction; the prediction models constructed with K-mer length from 3 to 5 were validated through a 10-fold cross-validation and independent testing dataset (Supplementary Tables 9–14, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The result shows that the prediction model based on 4-mer outperformed models based on 3-mer and 5-mer on both inter- and intravalidation, with an ROC-AUC of 0.860 on 10-fold cross-validation at the strictest genus level and an ROC-AUC of 0.801 on independence test dataset at the strictest genus level (Supplementary Tables 10 and 13, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Meanwhile, the ROC-AUC of the model based on 3-mers shows 0.853 on 10-fold cross-validation of the training set and shows 0.797 on the independent test set (Supplementary Tables 9 and 12, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). In addition, the model constructed with 5-mer gave an ROC-AUC of 0.841 on 10-fold cross-validation of the training set at the genus level and shows that of 0.785 on the test set (Supplementary Tables 11 and 14, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Therefore, the local K-mer descriptor was designed with a WL of 9000 bp and SS of 1800 bp and K-mer length of 4, and we constructed a phage–bacteria interaction prediction model based on different machine learning algorithms such as Random Forest, XGBoost, SVM and deep learning algorithms such as MLP, CNN, RNN and Bi-RNN. The performance of the

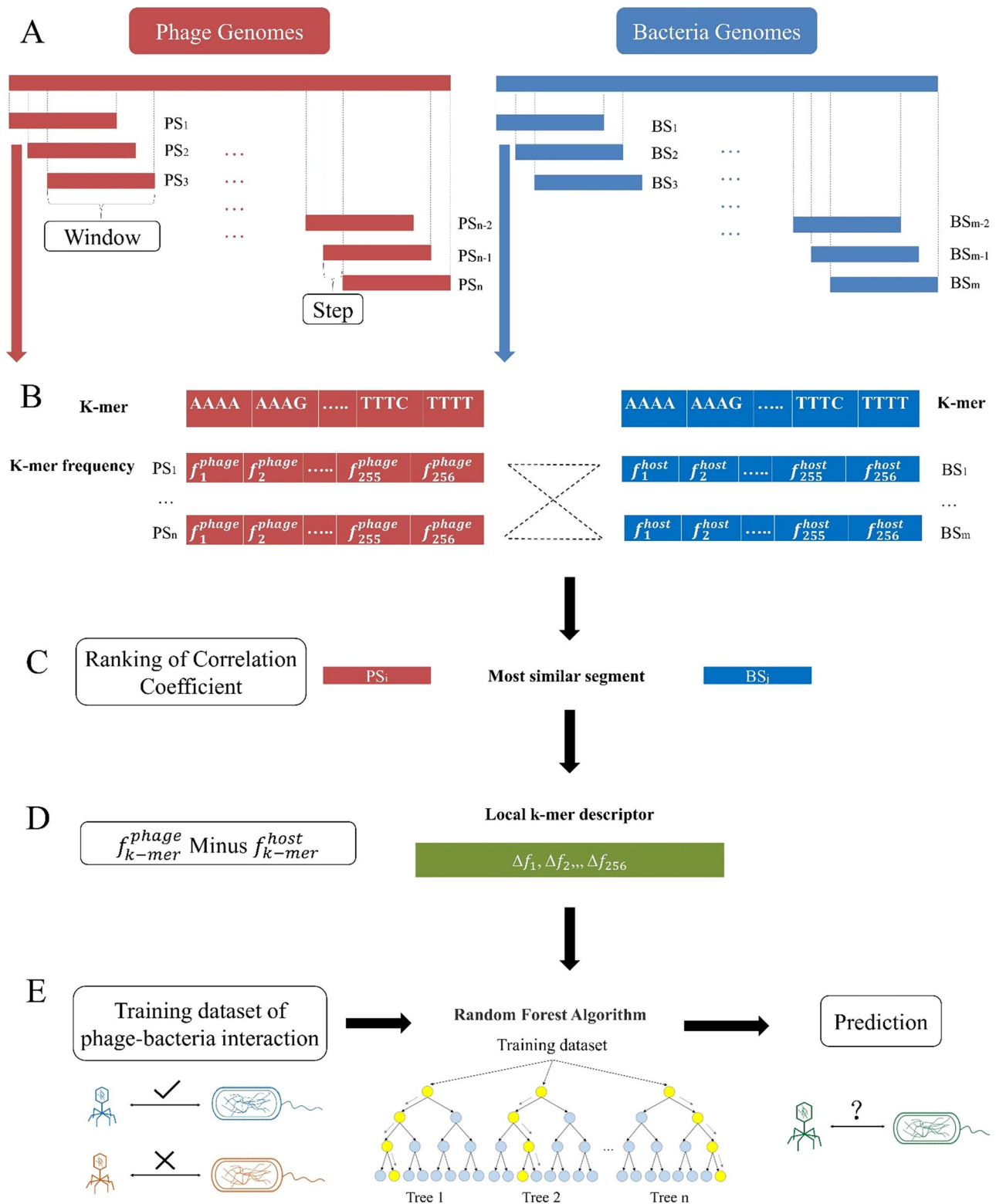


Figure 1. Workflow of PB-LKS approach. (A) Dividing the phage genome and the bacteria genome into multiple segments with defaulted WL and SS. (B) Generating the K-mer profile of all the phage–bacteria segment pairs. (C) Ranking the pairwise segments based on the correlation coefficient of K-mer descriptors. (D) Defining the model descriptor as the difference between the K-mer feature from phage (f_{k-mer}^{phage}) and the K-mer feature from host bacteria (f_{k-mer}^{host}). (E) Modeling the phage–bacteria interaction through the optimal algorithm.

above models is evaluated through 10-fold cross-validation on the training set and further test on independence test set at different taxonomy levels including kingdom, phylum, class, order, family and genus (Supplementary Tables 15–28, see Supplementary

Data available online at <http://bib.oxfordjournals.org/>). For 10-fold cross-validation, prediction models constructed by Random Forest, XGBoost and SVM show an ROC-AUC of 0.860, 0.864 and 0.870 at the genus level, while models constructed by MLP, CNN,

Table 1: Performance of machine learning approaches on initial testing by setting the WL of 1000 bp and SS of 1000 bp

Machine learning approaches	Accuracy	Sensitivity	Specificity	Precision	F-score	ROC-AUC
Bayesnet	0.606	0.606	0.606	0.622	0.593	0.695
Hoeffding Tree	0.613	0.613	0.613	0.622	0.605	0.698
Logistic Regression	0.712	0.712	0.712	0.712	0.712	0.780
Random Tree	0.688	0.688	0.688	0.688	0.688	0.688
Random Forest	0.727	0.715	0.739	0.720	0.713	0.803

RNN and Bi-RNN give an ROC-AUC value of 0.782, 0.782, 0.770 and 0.717, respectively (Figure 2A and Supplementary Tables 15–21, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Meanwhile, the Random Forest model and XGBoost model show an ROC-AUC of 0.801 and 0.817 at the genus level on the independence dataset. Compared with RF and XGBoost, the SVM model and deep learning models illustrated relatively lower prediction performance; the ROC-AUC value of four deep learning models is less than 0.8 at the genus level (Figure 2B and Supplementary Tables 22–28, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The 10-fold cross-validation result illustrated the superior performance of machine learning models, and the worse performance of DL methods on the independence test set further indicated that deep learning approaches are unsuitable for the PB-LKS model based on the local searching strategy K-mer descriptors. Meanwhile, the model constructed by XGBoost could achieve Minor enhancements than those of RF on all the tested scenarios with a difference of less than 2.5% at all taxonomy levels (Supplementary Tables 15–16, 22 and 23, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Considering the interpretability of the different models, we finally chose Random Forest to construct the PB-LKS model but also retained the Python package constructed by XGBoost for potential usage.

High performance of PB-LKS on general phage–bacteria interaction prediction

With the optimized parameters and algorithm, the PB-LKS was evaluated through 10-fold cross-validation at different taxonomy levels including kingdom, phylum, class, order, family and genus. Common evaluation parameters for binary classification including ROC-AUC, accuracy, recall, specificity, precision and F-score were used for model validation (see ‘Supplementary Model construction and evaluation’ part, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). It can be found that from the kingdom level to the genus level, the ROC-AUC value decreased from 0.959 to 0.860 (Figure 3A and Supplementary Table 15, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). This is reasonable that the prediction performance decreased with the increasing classification fineness because, for a broader taxonomy level, it is easier to distinguish the difference between positive pairs from negative ones. More importantly, even in the most restrictive prediction at the genus level, the PB-LKS could provide a high prediction performance with an ROC-AUC of 0.860, an accuracy of 0.775 and a recall of 0.792 (Supplementary Table 15, see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

Furthermore, for independent testing, the performance was relatively stable, with ROC-AUC value ranging from 0.752 to 0.852 (Figure 3B), which illustrated that the computational model could provide accurate prediction at different taxonomy levels. The prediction at the taxonomy level of phylum could achieve the

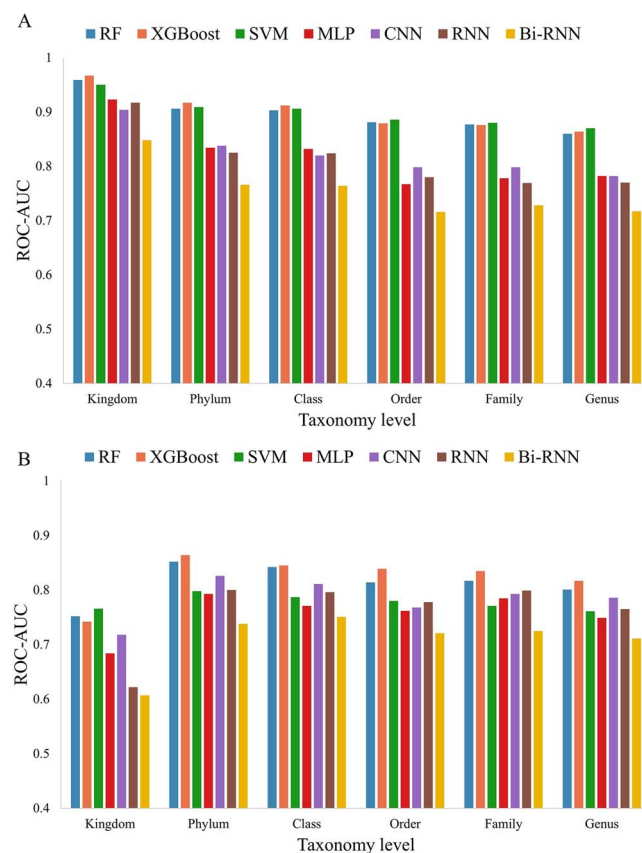


Figure 2. Performance of models constructed by different machine learning algorithms and deep learning algorithms at different taxonomy levels. (A) Ten-fold cross-validation performance on the training set. (B) Validation performance on the independent test set.

best prediction performance with an ROC-AUC value of 0.852 and followed by the taxonomy level of class with an ROC-AUC value of 0.842 (Figure 3B and Supplementary Table 22, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The performance at the genus level could reach the ROC-AUC of 0.801 and accuracy of 0.801. Moreover, considering the potential imbalanced situation in reality that positive pairs are expected to be significantly less than negative ones, while it is more important to discover positive samples than negative ones, we introduced PR-AUC for evaluation. Results showed that the PR-AUC could obtain performance from 0.848 to 0.931 at different taxonomy levels (Figure 3C). Similarly, the taxonomy of phylum achieved the best prediction performance with a PR-AUC value of 0.931 and followed by the taxonomy level of class with a PR-AUC value of 0.924 (Figure 3C). For the strictest genus level, PB-LKS could achieve the PR-AUC of 0.892. All above illustrated that PB-LKS could provide a good prediction performance of the phage–bacteria interaction based on multiple perspectives of evaluation, indicating the ability to provide an accurate determination of

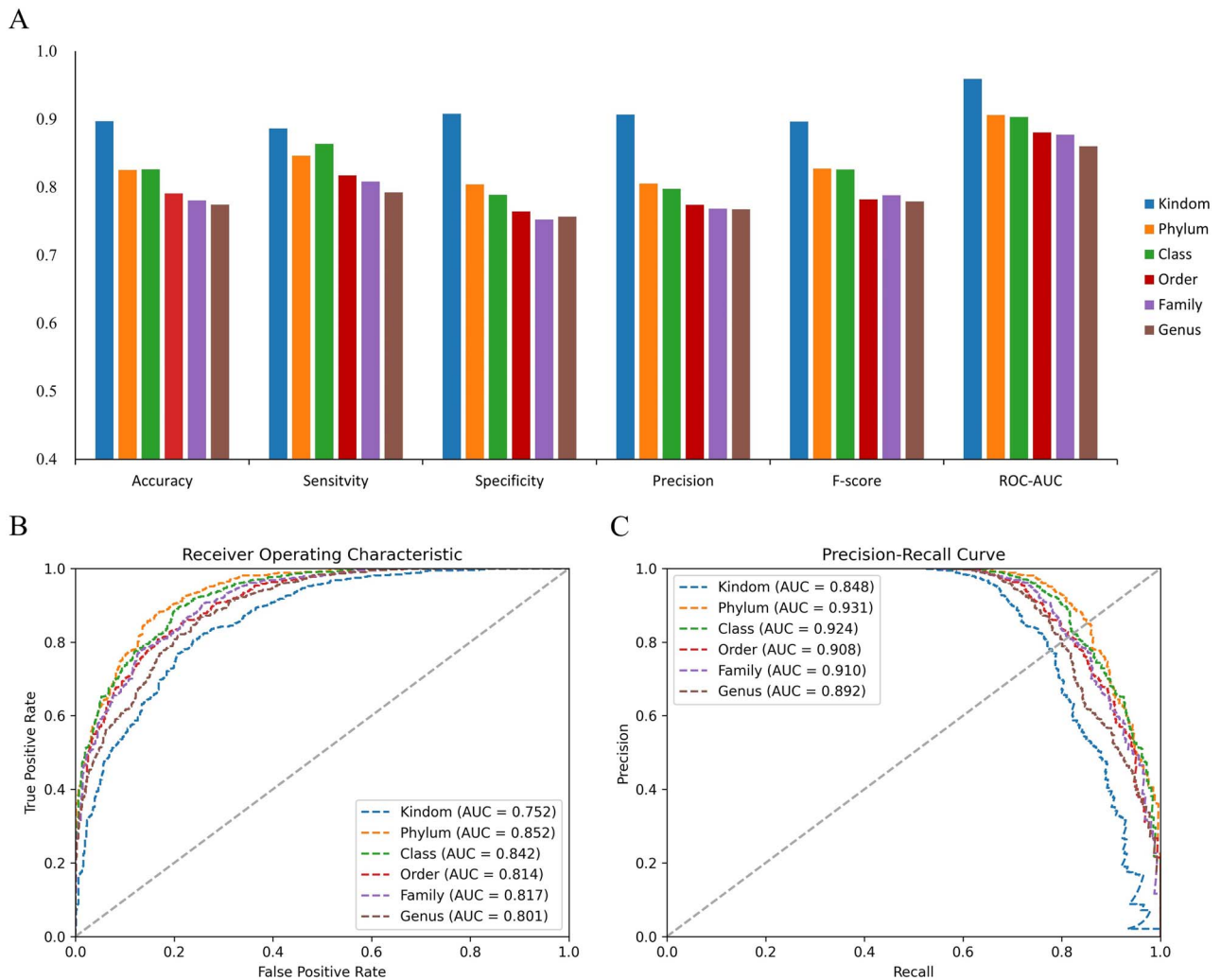


Figure 3. Model performance of constructed Random Forest model at different taxonomy levels. (A) Ten-fold cross-validation performance. (B) The Receiver Operating Characteristic curve on the independent test dataset. (C) The Precision-Recall curve on the independent test dataset.

Table 2: Performance of PB-LKS model and other state-of-the-art methods at the genus levels on independent test set

Predicted method	Accuracy	Recall	Specificity	Precision	F-score	ROC-AUC
PB-LKS	0.801	0.801	0.801	0.803	0.801	0.801
PHP	0.730	0.730	0.729	0.813	0.710	0.730
BLAST	0.751	0.751	0.751	0.829	0.736	0.751
CRISPR	0.546	0.545	0.546	0.762	0.427	0.546

phage–bacteria interactions at different taxonomy levels. Note that, the predicted probability score ranged between 0 and 1. The 267 wrong predictions included 108 false negatives and 159 false positives. Further investigation showed that 20% of the scoring region (0.4, 0.6) contains 54.682% of the wrong predictions, which means that this is the ‘fuzzy region’ that requires extra attention.

Finally, we compared PB-LKS to state-of-the-art alignment-free and alignment-based models. Considering the potential utility of PB-LKS in phage therapy, correctly predicting ‘positive samples’ is regarded as the most important function, recall was therefore an important measurement to evaluate the performance of these models.

The prediction performance on the same testing dataset of 671 phage–bacteria pairs illustrated that the alignment-free model PHP [23] based on K-mer frequencies could only provide a recall rate of 0.46 at the genus level, which is defined as host prediction

accuracy in other researches. For the alignment-based model, CRISPR [29] and BLAST [30] algorithms could achieve a recall of 0.59 and 0.77 at the genus level, respectively. Furthermore, we also testified the PB-LKS model with state-of-the-art methods of PHP, BLAST and CRISPR on the independence test set. Results showed that, at the genus level, above state-of-the-art methods could be outperformed by PB-LKS with an ROC-AUC of 0.80 and a recall of 0.80 (Table 2). Meanwhile, other methods showed ROC-AUC from 0.546 to 0.730 at the genus level, and the BLAST and CRISPR methods can only predict hosts for 596 and 89 phages from the independence dataset which includes 671 phages. Detailed comparison results at the different taxonomy levels are listed in Supplementary Tables 29–31 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>). According to these results, the PB-LKS model could conduct a comparable prediction performance at the strictest genus level compared with the

Table 3: Pickup rate for prediction of Actinobacteriophages–host interaction at different taxonomic levels

Host Strain	Number of Phages	Class	Order	Family	Genus
<i>M. smegmatis</i> mc ² 155	1739	99.655%	99.540%	98.735%	98.735%
<i>G. terrae</i> 3612	418	93.301%	93.541%	93.541%	92.823%
<i>M. foliorum</i> NRRL B-24224	366	98.361%	98.361%	96.721%	97.268%
<i>Arthrobacter</i> sp. ATCC 21022	213	79.812%	76.995%	76.526%	77.934%
<i>S. griseus</i> ATCC 10137	105	91.429%	84.762%	84.762%	89.524%

Table 4: Prediction result of PB-LKS and other state-of-the-art tools and experimental result of interaction between phage ϕ Ab124 and bacteria B1–B9 at different taxonomic levels^a

	B1	B2	B3	B4	B5	B6	B7	B8	B9
Experimental result	1	0	1	1	0	0	1	1	1
Predicted label of BLAST	0	0	0	0	0	0	0	0	0
Predicted label of PHP	0	0	0	0	0	0	0	0	0
Predicted label of PB-LKS	1	1	1	0	1	0	1	1	1
Predicted Score of PB-LKS	0.519	0.564	0.519	0.449	0.521	0.489	0.524	0.627	0.625

^a'1' represents that the interaction of phage ϕ Ab124 and the bacteria was predicted as phage–host, '0' represents that the interaction of phage ϕ Ab124 and the bacteria was predicted as phage–nonhost.

alignment-based models and with the wide application scope as the alignment-free methods.

High performance at the class level of Antinobacteriophage–host interaction

Actinobacteria is a group of bacteria that exhibits a cosmopolitan distribution [31], which can cause a variety of bacterial infection-based diseases, such as tuberculosis [32], leprosy [33] and actinomycosis [34]. To further verify the effectiveness of PB-LKS to identify the potential therapeutic phages, which target the bacteria that cause the above diseases, we narrow down the prediction spectrum from the kingdom level to a specific order of Actinobacteria. The independent testing dataset was derived from PhagesDB [27], which included 3455 pairs of Actinobacteriophage–host interactions. Here, we re-evaluate the above experimentally identified pairs through PB-LKS to calculate the pick-up rate for different taxonomies. Results showed that the pick-up rate could be reached at 92.851% (genus), 93.227% (family), 93.777% (order) and 94.588% (class), respectively, indicating the good prediction performance of the PB-LKS model.

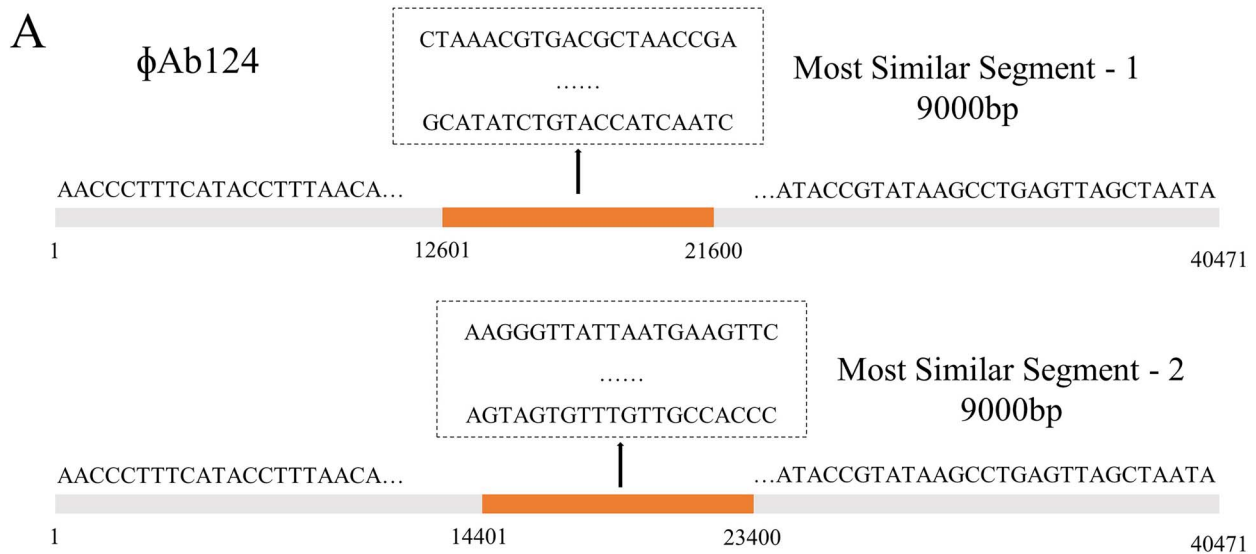
Furthermore, the host of 3455 Actinobacteriophages includes 67 bacterial strains, all of which belong to the Actinomycetes at the class level. Among them, the top five strains contained over 80% of the phage–bacteria pairs, including *Mycobacterium smegmatis* mc²155, *Gordonia terrae* 3612, *Microbacterium foliorum* NRRL B-24224, *Arthrobacter* sp. ATCC 21022 and *Streptomyces griseus* ATCC 10137 (Supplementary Figure 1, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The prediction performance on these strains was systemically evaluated through PB-LKS (Table 3). Note that, for the most abundant strain of *M. smegmatis* mc²155, 1739 phage–bacteria pairs were included and the pick-up rate for all taxonomy could reach a high level of 98.735%. Similar good performance can be found for the second and third abundant groups of *G. terrae* 3612 and *M. foliorum* NRRL B-24224, while the pick-up rate could reach 92.823% and 97.268% at the strictest level of genus, respectively. It is also noted that the prediction performance was different even on different strains, the pick-up rate reduced to 77.934% and 89.524% when the data abundance started to decrease. The above results showed that PB-LKS could achieve good prediction performance at the class taxonomy level, the reduced performance on *Arthrobacter* sp. ATCC 21022 and *S. griseus* ATCC 10137 might be caused by the following

possible reasons: (i) the biases due to the small amount of data and (ii) the strains rarely involved in the PB-LKS model (only two pairs for *Arthrobacter* available in our dataset). Meanwhile, the results indicated that further improvements were also needed when applying the PB-LKS to specific cases.

The potential utility of PB-LKS for clinical practice

The phage therapies used in clinical trials usually need to select the phage that could lyse the specific bacteria strain, which requires the accurate prediction of phage–bacteria interaction at the strain level. Thus, we further evaluate the prediction of PB-LKS in strain level to testify the potential utility in clinical usage. Here, the nine tested *A. baumannii* bacteria were derived from CRAB-infected patients or *in vitro* evolution, which were labeled as B1–B9 (see 'Supplementary Methods Bacteria Strain isolation and sequencing' part, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), respectively. Whether previously reported phage ϕ Ab124 could invade and lyse these nine bacteria were evaluated through PB-LKS, which has been experimentally verified (see 'Supplementary Methods Phage susceptibility assay' part, see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

We first tested the performance of current available alignment-free or alignment-based methods on the above phage and bacteria. Results showed that the alignment-free method of PHP [23] predicted the host of ϕ Ab124 as *Moritella* sp. PE36, which belongs to the taxonomy level of Bacteria, Pseudomonadota, Gammaproteobacteria, Alteromonadales, Moritellaceae and *Moritella*. However, B1–B9 are identified as *A. baumannii*, which belongs to the taxonomy level of Bacteria, Pseudomonadota, Gammaproteobacteria, Moraxellales, Moraxellaceae and *Acinetobacter*. Thus, the PHP could only provide correct prediction at the class level. On the other hand, the alignment-based method of BLAST can correctly detect the host of ϕ Ab124 at the species level by predicting it as *A. baumannii* strain AB179-VUB (Supplementary Table 32, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). These two methods give a prediction accuracy of 33.3% for the interactions between phage ϕ Ab124 and bacteria B1–B9, which predicted all nine pairs as negative (Table 4). This means that the alignment-based method could provide accurate prediction at the species level but fail to provide the specific host strains.

**B**

	B1	B2	B3	B4	B5	B6	B7	B8	B9
Information of strain	LDB8040	LDB8040- ϕ Ab124R	SGH7143	TJG6288	WHC8200	WHC10114	WHC5364	WHC5408	JJM1537
			P3 original	P1 original	P4 in vivo evolved	P4-10114	P4 original urine	P4 original sputum	
ϕ Ab124	S		S	S			S	S	S

Susceptible Resistance

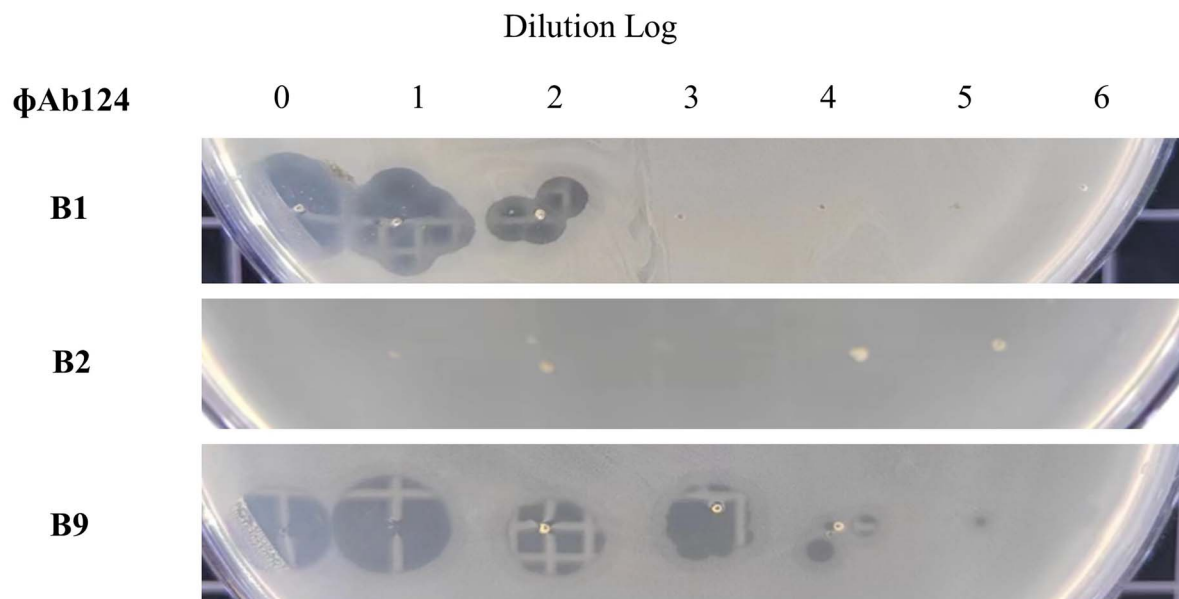
C

Figure 4. (A) Two MSSs of phage ϕ Ab124 screened by PB-LKS between phage ϕ Ab124 and bacteria B1–B9. (B) Phage susceptible pattern of nice CRAB strains to phage ϕ Ab124. P1–P4 representing the strains derived from patient 1 to patient 4 in our previous study [28]. (C) The Efficiency of Plating assay of phage ϕ Ab124 on bacteria B1, B2 and B9.

Meanwhile, PB-LKS could successfully predict that the *A. baumannii* is the host for ϕ Ab124. Interaction prediction between phage ϕ Ab124 and bacteria B1–B9 by PB-LKS model was based on the screened MSSs. The prediction result shows that 2 MSSs on

ϕ Ab124 were detected, one in the range from 12 601 to 21 600 for B7, while the other MSS in the range from 14 401 to 23 400 for B1, B2, B3, B4, B5, B6, B8 and B9 (Figure 4A). Furthermore, the predicted scores of PB-LKS illustrated that bacteria B1, B2, B3, B5, B7, B8 and

Table 5: Feature importance of screened core features at different taxonomy levels

Core features	Kingdom	Phylum	Class	Order	Family	Genus
TAAA	0.017	0.020	0.032	0.019	0.023	0.013
TTAA	0.017	0.015	0.024	0.020	0.020	0.025
TTTA	0.015	0.020	0.024	0.020	0.016	0.017
ATAT	0.014	0.007	0.013	0.011	0.008	0.007
AAAT	0.014	0.013	0.016	0.012	0.017	0.011
GTCG	0.011	0.023	0.017	0.008	0.017	0.014
AATA	0.008	0.008	0.008	0.014	0.010	0.009
TCGA	0.008	0.010	0.009	0.007	0.009	0.005
TTAT	0.008	0.027	0.014	0.016	0.015	0.011
ATTA	0.008	0.013	0.017	0.024	0.013	0.019

B9 were over the threshold (>0.5), while B4 and B6 were below the threshold (Table 4), which indicates that seven strains were the potential host for ϕ Ab124. The phage lysis experiments indicated that B1, B3, B4, B7, B8 and B9 were sensitive for phage ϕ Ab124, while B2, B5 and B6 were insensitive (Figure 4B). Besides B3–B8 that were validated before [28], bacterial plaque of three new strains including B1, B2 and B9 were further tested through the efficiency of plating assay (Figure 4C). The experiments showed that PB-LKS could give an accuracy of 66.667% (6/9) on all tested samples, a recall rate of 83.333% (5/6) on all positive samples and a precision of 71.429% (5/7).

Moreover, the predicted probability score of PB-LKS between ϕ Ab124 and nine bacteria ranged from 0.449 to 0.627, while the three mispredictions were B2 (score of 0.564), B4 (score of 0.449) and B5 (score of 0.521), which were located on the fuzzy region (Table 4). This result is consistent with our independent testing, while the performance of PB-LKS would be reduced on the fuzzy region (see Result part ‘High performance of PB-LKS on general phage–bacteria interaction prediction’). On the other hand, the correct prediction of B8 and B9 could achieve scores of 0.627 and 0.625, respectively. This indicates the good performance of the model outside the fuzzy region. Among the tested bacteria, the pair of ϕ Ab124–B8 pair obtained the highest score, indicating the possible interaction. Moreover, the ϕ Ab124-based therapy was proven with effective lytic activity *in vivo* for the treatment of COVID-19 patients with secondary *A. baumannii* infection of B8 [28], which indicated the potential clinical usage of PB-LKS to design the pre-optimized phage therapy.

Detecting the important local K-mers for phage–bacteria interaction

Furthermore, the feature importance of all 256 K-mer features was screened to detect the important K-mer features, which represented the core local K-mers that play essential roles in the phage–bacteria interactions and contributed to the model prediction. This involved two steps: (i) calculate the statistical significance of each core K-mer and (ii) calculate the feature importance of each core K-mer at different taxonomy levels (see ‘Supplementary Methods Core feature screening’ part, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Considering the interpretability, the feature importance of the Random Forest classifier was derived for further analysis.

For step i, the essential cores with statistically significant scores (P value <0.001) between positive and negative samples were derived, which included 159 core K-mers (Supplementary Table 33, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). For step ii, the feature importance of each core K-mer at different levels of kingdom

(Supplementary Table 34, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), phylum (Supplementary Table 35, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), class (Supplementary Table 36, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), order (Supplementary Table 37, see Supplementary Data available online at <http://bib.oxfordjournals.org/>), family (Supplementary Table 38, see Supplementary Data available online at <http://bib.oxfordjournals.org/>) and genus (Supplementary Table 39, see Supplementary Data available online at <http://bib.oxfordjournals.org/>) were evaluated through sklearn with the defaulted threshold. Here, 11 overlapped features among all 6 taxonomy levels were derived, involving 10 statistically significant core K-mers of TAAA, TTAA, TTTA, ATAT, AAAT, GTCG, AATA, TCGA, TTAT and ATTA. As illustrated in Table 5, the feature importance of 10 essential core K-mers ranged from 0.005 to 0.032, 1.28–8.19 times than the average feature importance of 3.9×10^{-3} (1/256).

Moreover, according to the assumptions of PB-LKS, the important local core K-mers should illustrate similar distributions among the compared genome segments from the phage and the bacteria in positive samples, while should illustrate random distribution in negative tones. To test this assumption, the frequency distribution of 12 core K-mers including the above 10 important ones, and K-mers ranked the last two among 256 K-mers in our dataset were evaluated.

As illustrated in Figure 5 and Supplementary Figure 2, see Supplementary Data available online at <http://bib.oxfordjournals.org/>, the X-axis illustrated the frequency of corresponding core K-mers in the phage genome, and the Y-axis illustrated the frequency in the bacteria genome. Results showed that the Pearson correlation coefficient (Pcc) of important local core K-mers between phage and bacteria genome could reach 0.777–0.897. The core K-mers TAAA could obtain the highest Pcc of 0.897, which demonstrated the consistent distribution of local similar genome sequences for phage and bacteria. In negative samples, the Pcc of all 10 important core K-mers ranged from 0.105 to 0.434, illustrating relatively low and random distribution between two compared genomes (Figure 5A–H and Supplementary Figure 2A–L, see Supplementary Data available online at <http://bib.oxfordjournals.org/>). In contrast, core K-mers with low contributions illustrated relatively low correlations in both the positive group and the negative group. For example, the non-core K-mers TGGA and CATG with the lowest feature importance among all 256 K-mers showed that the Pcc of these two K-mers is 0.3711 and 0.2021 in positive samples, while that in negative samples is 0.1340 and 0.0247 (Figure 5I–L). The Pcc difference of non-core K-mer between positive and negative samples is smaller than that of

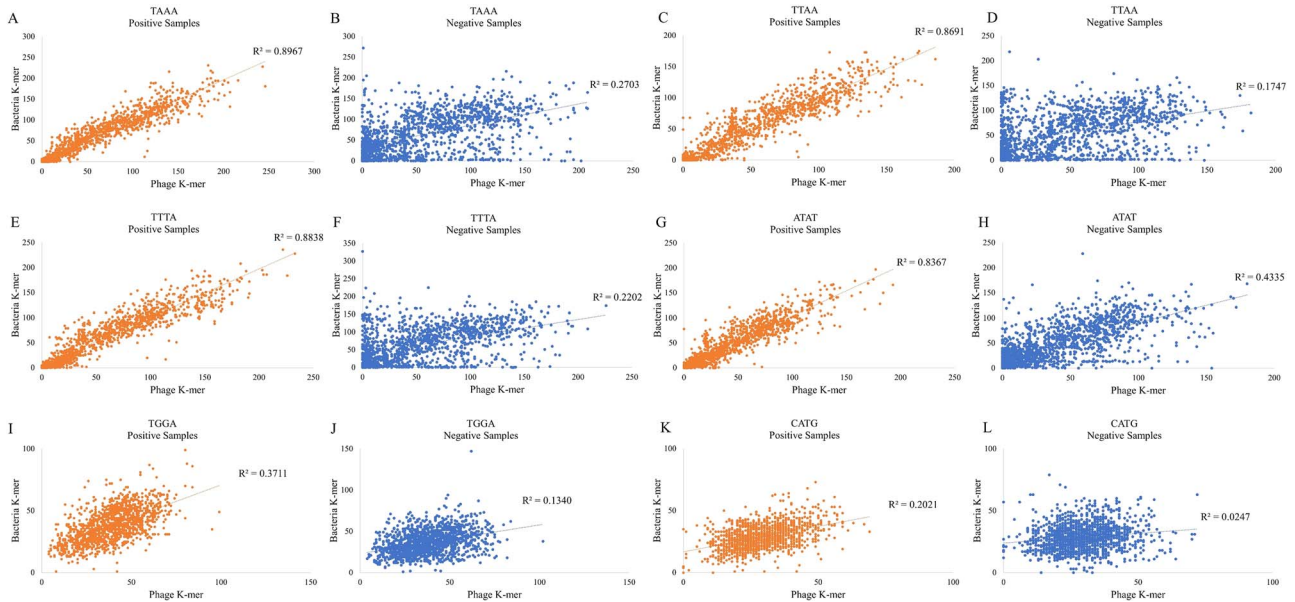


Figure 5. The frequency distribution of six K-mers between viral and bacterial genomes on positive samples and negative samples. (A) Frequency distribution of TAAA among positive samples. (B) Frequency distribution of TAAA among negative samples. (C) Frequency distribution of TTAA among positive samples. (D) Frequency distribution of TTAA among negative samples. (E) Frequency distribution of TTTA among positive samples. (F) Frequency distribution of TTTA among negative samples. (G) Frequency distribution of ATAT among positive samples. (H) Frequency distribution of ATAT among negative samples. (I) Frequency distribution of TGGA among positive samples. (J) Frequency distribution of TGGA among negative samples. (K) Frequency distribution of CATG among positive samples. (L) Frequency distribution of CATG among negative samples. The Pearson coefficient R^2 was calculated for K-Mer frequency in genomes of phage and bacteria.

core K-mer, which implies that the distribution of nonsignificant features on compared local segments is less relevant between phage and bacteria.

Model implementation

Workflow of PB-LKS

The package of PB-LKS accepts nucleotide sequence data of both phage and bacteria as input and outputs the predicted phage–host relationship classification between input phage and bacteria at the genus level. The workflow of PB-LKS included the following three steps.

(1) Data inputting. The PB-LKS accepts *.fasta* format file of phage and bacterial genome as input. The examples of input format were provided in the ‘Example’ folder of the Python package, and both the file of *Bacteria genome.fasta* and the file of *Phage genome.fasta* were needed.

(2) Descriptor generating. After inputting the genome sequences of phages and bacteria, PB-LKS will automatically split the nucleotide sequences into small segments with optimized WL and SS. Then, the MSS pair was derived and the K-mer-based descriptors were generated.

(3) Prediction outputting. Finally, PB-LKS used the well-trained Random Forest classifier to predict the phage–host relationship between the input two genomes.

Worked example

The example case illustrated the prediction of the relationship between *Aeromonas* phage phiO18P and *Escherichia coli* STEC_94C through PB-LKS. The genome of *Aeromonas* phage phiO18P contained 33.985 kbp, and the *.fasta* file was downloaded from the NCBI RefSeq database (Accession ID: DQ674738). The genome *.fasta* file was named ‘*Phage genome.fasta*’ under the file folder of *Example*, with a file size of 34.1 KB. *Escherichia coli* STEC_94C (Accession ID: GCF_000225105.1) is the host bacteria with a genome size

of 5 Mbp, and the *.fasta* file was named ‘*Bacteria genome.fasta*’ under the file folder of *Example*, with a file size of 4.85 MB. According to the README.md file provided at the website of <https://github.com/wanchunnie/PB-LKS>, we need to store the PB-LKS package on the local drive (for example, C:/) before prediction, and the Scikit-learn package [35], Biopython package [36] and Numpy package [37] are required to run the PB-LKS package, which can be installed by invoking `pip install`. The usage is displayed as follows:

```
# Local environment setup
1. install conda(to manage environment)
2. Change directory to the path of this project
'''bash
cd {your_path_to_PB-LKS}
'''
3. Run the following codes in your terminal
'''bash
conda create -n PB-LKS python=3.10
conda deactivate (if base environment is activated)
conda activate PB-LKS pip install -r requirements.txt
'''

# Usage of PB-LKS
1. Command-line interface
'''bash
cd {your_path_to_PBLKS}
# display help message of PB-LKS CLI
python PB-LKS.py -h
'''
2. Run the PB-LKS with the example
'''bash
python PB-LKS.py -p {your_path_to_phage's fasta file} -b {your_path_to_bacterial fasta file} -ba -o {folder to save result file}
'''

# Detailed arguments mentioned above and their usages are listed as follows
options:
-h, --help show this help message and exit
-p PHAGE, --phage PHAGE path/folder to your phage sequence file(in fasta format)
-b BAC, --bac BAC path/folder to your bacteria sequence file(in fasta format)
-ba, --batch run prediction in batch, -p,-b should be folder if set True
-o OUTPUT, --output OUTPUT the folder you want to save batch prediction results (results will be printed into the terminal by default)
```

Table 6: The strictest prediction taxonomy level of current phage–host interaction prediction tools

Prediction method	Prediction level	Prediction method	Prediction level
VirHostMatcher [20]	Genus	HostPhinder [19]	Species
WisH [24]	Genus	LMFH-VH [21]	Species
PHP [23]	Genus	ILMF-VH [22]	Species

DISCUSSION

Phages could invade and lyse bacteria while being harmless to mammalian cells. Thus, it has a wide range of application scopes including identifying bacteria, lysing the food-borne pathogenic bacteria in the production environment during food processing and serving as a therapy for drug-resistant bacteria. Determination of the phage–bacteria relationship was important in the application and screening of phage. For *in-silico* prediction of phage–bacteria interaction, the main idea of the algorithm design is to detect those potential integrated phage gene segments from the bacteria genome. Previous studies have demonstrated oligonucleotide frequency as an obvious signal to identify phage–host interaction [20]. Also, it is worth noting that, besides genome integration, phage can drive bacterial evolution through high selective pressures, lysogenic conversion, transduction and host gene disruption [38], while those usually occur on the partial sequence segment of the genome rather than on the whole genome sequence. In addition, current tools based on genome analysis can only predict the host at the genus or species level (Table 6). Nevertheless, phage therapy requires the recognition of interactions between therapeutic phages and specific bacteria strains. Thus, we designed the local K-mer strategy to achieve phage–host interaction prediction, which aimed to determine the local similarity of genomes between phages and bacteria, and has better potential in identifying K-mer dissimilarity between phages and mutant bacteria at the same species level with host bacteria. All the possible sequence segments were first split from the whole genome sequence of both phages and bacteria. Then, the MSS pair was selected for K-mer scoring construction. To obtain the optimized WL and SS, different parameter combinations were evaluated to make the balance between the prediction performance and time complexity. Finally, PB-LKS could predict the relationship between specific phage and bacteria mutants, which illustrated potential clinical usage to design phage-related therapies.

Notably, the major difficulty in model construction is the design of a negative dataset. The currently released dataset often contains accurate ‘true positive’ samples that were validated by the experimental to ensure that the tested positive bacteria is the host for the specific phage [27, 39]. Meanwhile, it is not certain that an un-tested bacteria is not the host for any phages. In short, the absence of evidence is not the same as evidence of absence. To address that, for negative dataset selection, we carefully choose those bacteria that are different from the host bacteria at the genus level. In other words, we excluded all the bacteria within the same genus level of all validated host bacteria. Considering the antibacterial spectrum of bacteriophages is often extremely narrow, with few cross-reaction cases [40, 41], which means that most bacteriophages are often only useful for one kind of bacteria. On the other hand, it will be more accurate to expand the selection of ‘true negative’ from the genus level to the family level or even the order level, in which the negative ones will be more ‘negative’. However, it is not good for feature extraction and model

construction, because the difference between bacteria from two Families will be too big, and it is difficult to accurately detect the K-mer features that could distinguish positive and negative samples. In addition, considering that the whole background dataset contains 31 918 bacteria, which is exceedingly larger than the 671 positive sample set, the potential positive data contained therein would be very sparse after excluding all bacteria from the same genus level of the positive bacteria and selecting 671 bacteria from the background dataset containing over 30 000 bacteria. More importantly, for phage therapy, finding the phages that could interact with pathogenic bacteria is the key. For this purpose, the missing potential ‘true positive’ is not as important as ensuring the predicted ‘positive samples’ are correct. Based on the above reasons, we think that the current way to select negatives is suitable for model construction.

In addition to the dataset, selecting a suitable model is also important for the construction of the PB-LKS model. Here, we tested several widely used machine learning models and deep learning models that used to predict genome or protein interactions [42]. It is testified that the performance of the DL-based model is inferior to that of the tree-based model. Furthermore, DL models with complex architecture such as CNN and RNN exhibit worse performance compared with simpler ones such as Vanilla-MLP. Those phenomena may be caused by the following reasons: (i) tree-based models are more suitable for tabular data than deep learning approaches because of specific features for tabular data [43, 44]. (ii) The limited size of the training dataset (2852 pairs of phage–bacteria interactions) compared with the relatively larger length of the LKS-based descriptor. A longer feature vector implies a larger number of parameters in the model, while empirical evidence and literature suggest that the number of samples used to train a deep-learning model should be significantly larger than the number of its parameters [45]. (iii) Complicated architectures may not always bring better performance if the form of the training set and the characteristics of the model do not match. Advanced models such as CNN and RNN are specially designed to capture temporal or spatial relationships in tasks such as natural language processing and computer vision [46–48]. However, the values in the LKS-based descriptor of this work are exclusively handcrafted features, and there is no such relationship between features. In that case, a tree-based classifier might be more suitable for the Local K-mer strategy, which generates tabular data as features. Meanwhile, for the tree-based classifier used in this study, Random Forest is an integrated learning method which consists of multiple decision trees and shows good interpretability [49]. Each decision tree is trained independently and the subsequent prediction is based on the average or majority vote of all the decision trees. This process makes it easy to understand by looking at the information such as decision paths, feature importance and split point of each decision tree. However, XGBoost uses gradient boosting techniques to improve the accuracy; each new decision tree is constructed to make the residuals of the previous tree smaller, which leads to weak interpretability [50]. Considering the high prediction performance and more intuitive and concrete interpretability, to further detect the important K-mer for PB-LKS, Random Forest was selected as the optimal algorithm. In the meantime, the XGBoost-based PB-LKS was also provided in the Python package for users valuing the predictive accuracy over model interpretability. On the other hand, we noticed that other deep learning API such as Keras [51] may involve better selected layers, which hold the potential to give better prediction performance. It is worth trying multiple deep learning approaches based on different API to optimize the model of phage–bacteria

interaction prediction when more data are available in the future. In that case, we are planning to test multiple models based on Keras and create a user-friendly webserver of PB-LKS that integrates optimized deep-learning or machine-learning approaches in the future.

Compared with available alignment-free and alignment-based models, the PB-LKS model shows better performance with a recall of 0.80 on the independent testing dataset at the strictest genus level. Previous studies have demonstrated that the prediction accuracy of alignment-based algorithms is higher than that of alignment-free algorithms but with a narrower application scope [23]. Despite the above, PB-LKS is an alignment-free algorithm but can perform similar prediction accuracy as the alignment-based CRISPR model and significantly outperform the alignment-based BLAST model, which illustrates the advantages of PB-LKS over existing algorithms. The reliable performance of PB-LKS may take advantage of the step of comparing the correlation coefficient between local K-mer frequencies of phage and bacteria genome before constructing the K-mer features. Also, from feature importance analysis, we noticed that not all K-mer features contributed equally to the classification model. Among them, 10 nucleotide features including TAAA, TTAA, TTTA, ATAT, AAAT, GTCG, AATA, TCGA, TTAT and ATTA were the core K-mer features screened in our model. The analysis of those core features would further help us understand the interaction between phage and bacteria.

The comparison result of PB-LKS and other peer methods on clinical usage also proves the better performance of our model. PB-LKS shows 66.7% prediction accuracy at the strain level, while the BLAST-based method and PHP predict all nine bacteria strains as negative. In other words, these peer methods can only predict the candidate host for query phages instead of predicting interaction between phages and bacteria strains with gene mutation, which cannot be applied in clinical phage therapy against bacterial infection. Moreover, ϕ Ab124 illustrated good clinical utility in generating phage therapy on COVID-19 patients who suffered from secondary infection of CRAB [28]. Before phage therapy, several high-grade antibiotics had been applied to this patient but had failed to eliminate the bacteria, while the B8 isolated from this patient showed that ϕ Ab124 was the phage with effective lytic activity *in vivo* [28]. This result implies that the PB-LKS model has the potential to predict phage–bacteria interaction at the strictest criterion at the strain level, which testifies that PB-LKS could be applied for the development of clinical phage therapy.

Despite the good performance and potential utility, there are still several limitations for PB-LKS. At first, the prediction of a probability score around 0.5, which was defined as the fuzzy region, may not be correct. Both the high-throughput validation and experimental test indicated that the rate of incorrect predictions in the fuzzy region was increased than those in other regions. Second, the design principle of PB-LKS is that the phage and its host may share some describable sequence features rather than simple sequence similarity during their long history of symbiosis [14, 15]. In that case, special cases such as the phage-resistant bacteria inducted through the Next Evolution Phage-Typing strategy, such as B2 in here, may not be a good target for PB-LKS. Third, PB-LKS is a training-based model, and thus, there were the following factors, which might affect the performance of the *in-silico* model: (i) the accumulation of bacteria genome sequences and the detailed annotation of taxonomy classification (for example, the annotation of species for host bacteria) could help make more specific model. (ii) The accumulation of diverse determined phage–bacteria interaction as a training dataset could

further improve model performance. (iii) The integrity of genome sequencing data and the completed DNA fragment assembly could make it more accurate in screening the MSS through a sliding window.

Key Points

- The PB-LKS could predict the phage–bacteria relationship at the strain level and discern bacteria mutants, which could be useful to the development of phage therapy.
- The PB-LKS incorporates the local K-mer strategy, which focuses on the most similar segment detected among the whole genome to predict the phage–bacteria interaction. This strategy could accelerate the detection of antibacterial phages and pre-optimize the phage therapy for bacterial infection.
- Ten essential core K-mers were detected in PB-LKS that could successfully distinguish the positive interactions and negative ones. Those motifs contain the preference for phage to infect the bacteria and could help guide the design of functional phages.
- PB-LKS is applied on the pre-optimized phage therapy design for *A. baumannii* and illustrates better performance than the current state-of-the-art tools.
- The Python package for PB-LKS is freely available on GitHub (<https://github.com/wanchunnie/PB-LKS>).

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

Supplementary tables and figures. The information of training set, independent test set and case study set is freely available on GitHub (<https://github.com/wanchunnie/PB-LKS-dataset>). The codes and data to train and test the multiple models mentioned in this study are freely available on GitHub (<https://github.com/wanchunnie/PBLKS-TrainandTest>).

ACKNOWLEDGEMENTS

This work is supported by the Medical Science Data Center of Fudan University. We also thank the participating CRAB patients, and the health professionals who providing outstanding patient care at considerable personal risk, which are from Shanghai Public Health Clinical Center and Zhongshan Hospital.

FUNDING

National Natural Science Foundation of China (32000470, 32370697); National Key Research and Development Program of China (2022YFF1101104, 2021YFA0911200); Shanghai Commission of Science and Technology (20Y11900300).

AUTHORS' CONTRIBUTIONS

J.X.Q., W.C.N. and T.Y.Q. designed the model, constructed the computational model and wrote the manuscript. H.D. constructed and validated models based on deep learning methods. J.D. and N.N.W. perform the phage lysis experiments and bacteria sequencing. D.Z.L. modified the manuscript. Y.W.W., Y.X.Z., J.T.X. and X.X.T. validated the model. T.Y.Q. and N.N.W. co-supervised the whole project.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study using clinically sourced bacterial isolates was approved by the Ethic Committee of Shanghai Public Health Clinical Center on 28 March 2019, Approval No. 2017-S027-08.

DATA AVAILABILITY

The whole genome sequence data of nine *A. baumannii* bacteria are available in the NCBI Sequence Read Archive database, with accession number SRR24235333-SRR24235341.

REFERENCES

- Clokier MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. *Bacteriophage* 2011;**1**(1):31–45.
- Cisek AA, Dąbrowska I, Gregorczyk KP, Wyzewski Z. Phage therapy in bacterial infections treatment: one hundred years after the discovery of bacteriophages. *Curr Microbiol* 2017;**74**(2):277–83.
- 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.
- Lenski RE. Dynamics of Interactions between bacteria and virulent bacteriophage. In: Marshall KC (ed). *Advances in Microbial Ecology*, Vol. 10. Boston, MA: Springer, 1988, 1–44.
- Hanlon GW. Bacteriophages: an appraisal of their role in the treatment of bacterial infections. *Int J Antimicrob Agents* 2007;**30**(2):118–28.
- Oechslin F, Piccardi P, Mancini S, et al. Synergistic interaction between phage therapy and antibiotics clears pseudomonas aeruginosa infection in endocarditis and reduces virulence. *J Infect Dis* 2017;**215**(5):703–12.
- Middelboe M, Chan AM, Bertelsen SK. Isolation and life-cycle characterization of lytic viruses infecting heterotrophic bacteria and cyanobacteria. In: Wilhelm SW, Weinbauer MG, Suttle CA, (eds). *Manual of Aquatic Viral Ecology*, American Society of Limnology and Oceanography, Waco, TX, 2010, 118–33.
- Henry M, Biswas B, Vincent L, et al. Development of a high throughput assay for indirectly measuring phage growth using the OmniLog(TM) system. *Bacteriophage* 2012;**2**(3):159–67.
- Deng L, Ignacio-Espinoza JC, Gregory AC, et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* 2014;**513**(7517):242–5.
- Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* 2014;**15**(9):577–84.
- de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol* 2019;**27**(1):51–63.
- Edwards RA, McNair K, Faust K, et al. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 2016;**40**(2):258–72.
- Coclet C, Roux S. Global overview and major challenges of host prediction methods for uncultivated phages. *Curr Opin Virol* 2021;**49**:117–26.
- Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr Opin Microbiol* 2017;**38**:66–73.
- Jiang F, Doudna JA. CRISPR-Cas9 structures and mechanisms. *Annu Rev Biophys* 2017;**46**:505–29.
- Versoja CJ, Pfeifer SP. Computational prediction of bacteriophage host ranges. *Microorganisms* 2022;**10**(1):149.
- Carbone A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* 2008;**66**(3):210–23.
- Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* 2006;**7**:8.
- Villarreal J, Kleinheinz K, Jurtz V, et al. HostPhinder: a phage host prediction tool. *Viruses* 2016;**8**(5):116.
- 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.
- 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.
- 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.
- 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.
- 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.
- Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* 2020;**18**(3):125–38.
- Land M, Hauser L, Jun SR, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 2015;**15**(2):141–61.
- Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics* 2017;**33**(5):784–6.
- Wu N, Dai J, Guo M, et al. Pre-optimized phage therapy on secondary *Acinetobacter baumannii* infection in four critical COVID-19 patients. *Emerg Microbes Infect* 2021;**10**(1):612–8.
- 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.
- Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
- Mawang CI, Azman AS, Fuad ASM, Ahamad M. Actinobacteria: an eco-friendly and promising technology for the bioaugmentation of contaminants. *Biotechnol Rep (Amst)* 2021;**32**:e00679.
- Koch A, Mizrahi V. Mycobacterium tuberculosis. *Trends Microbiol* 2018;**26**(6):555–6.
- Mungroo MR, Khan NA, Siddiqui R. Mycobacterium leprae: pathogenesis, diagnosis, and treatment options. *Microb Pathog* 2020;**149**:104475.
- Stabrowski T, Chuard C. Actinomycosis. *Rev Med Suisse* 2019;**15**(666):1790–4.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
- Cock PJ, Antao T, Chang JT, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**(11):1422–3.
- Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature* 2020;**585**(7825):357–62.
- Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol* 2022;**20**(1):49–62.
- 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.
- 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.
41. Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brüßow H. Phage-host interaction: an ecological perspective. *J Bacteriol* 2004;**186**(12):3677–86.
 42. Horlacher M, Cantini G, Hesse J, et al. A systematic benchmark of machine learning methods for protein-RNA interaction prediction. *Brief Bioinform* 2023;**24**(5):bbad307.
 43. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? In: *Advances in Neural Information Processing Systems* 2022;**35**:507–20.
 44. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Information Fusion* 2022;**81**:84–90.
 45. Bengio Y Practical recommendations for gradient-based training of deep architectures. In: Montavon G, Orr GB, Müller KR, (eds). *Neural Networks: Tricks of the Trade*. Lecture Notes in Computer Science, vol **7700**. Springer, Berlin, Heidelberg, 2012, 437–78.
 46. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**(7553):436–44.
 47. Elman JL. Finding structure in time. *Cognit Sci* 1990;**14**(2): 179–211.
 48. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**(11):2278–324.
 49. Conesa A, Hernández R. Chapter 16 – Omics data integration in systems biology: methods and applications. In: García-Cañas V, Cifuentes A, Simó C, (eds). *Comprehensive Analytical Chemistry*. Elsevier, Amsterdam, Netherlands, 2014, 441–59.
 50. Chen T, Guestrin C. XGBoost: a scalable tree boosting system In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery: San Francisco, California, USA, 2016, p. 785–94.
 51. Chollet F. Keras: *The Python Deep Learning library*, Astrophysics Source Code Library, 2018. Available at: <https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract>.