Bioinformatics Advances, 2023, **00**, vbad101
https://doi.org/10.1093/bioadv/vbad101
Advance access publication 2 August 2023
Application Note

Sequence analysis

PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data

Jiayu Shang [1,†], Cheng Peng[1,†], Herui Liao [1], Xubo Tang [1], Yanni Sun [1*]

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong (SAR), China

*Corresponding author. Department of Electrical Engineering, Office G6411, YEUNG, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China. E-mail: yannisun@cityu.edu.hk

[†]Equal contribution.

Associate Editor: Sofia Forslund

Abstract

Motivation: There is accumulating evidence showing the important roles of bacteriophages (phages) in regulating the structure and functions of the microbiome. However, lacking an easy-to-use and integrated phage analysis software hampers microbiome-related research from incorporating phages in the analysis.

Results: In this work, we developed a web server, PhaBOX, which can comprehensively identify and analyze phage contigs in metagenomic data. It supports integrated phage analysis, including phage contig identification from the metagenomic assembly, lifestyle prediction, taxonomic classification, and host prediction. Instead of treating the algorithms as a black box, PhaBOX also supports visualization of the essential features for making predictions. The web server is designed with a user-friendly graphical interface that enables both informatics-trained and nonspecialist users to analyze phages in microbiome data with ease.

Availability and implementation: The web server of PhaBOX is available via: https://phage.ee.cityu.edu.hk. The source code of PhaBOX is available at: https://github.com/KennthShang/PhaBOX.

1 Introduction

As viruses that infect bacteria, bacteriophages (phages) are the most widely distributed and abundant biological entities in the biosphere (Cobián Güemes *et al.* 2016). With an estimated population of more than 10^{31} particles (Mushegian 2020), phages play an important role in modulating microbial system dynamics by lysing bacteria (Fernández *et al.* 2018). Recently, accumulating studies show that phages have an important impact on multiple applications, such as disease diagnostics (Roth *et al.* 2021), phage therapy (Petrovic Fabijan *et al.* 2020), and on the food industry (Cristobal-Cueto *et al.* 2021).

Although there are tools available for different tasks such as phage taxonomic classification and host prediction, these tools are often published as open-source codes. When the users want to characterize phages in metagenomic data, they need to install several tools and parse the intermediate files from different methods. However, installing some of these softwares requires informatics training and takes substantial computational resources. For example, some open-source codes are not accompanied with detailed user manuals, making the installation tedious and error-prone. Thus, it is preferred to have an integrated web server that can conduct comprehensive phage analysis while sparing the users the time and trouble of installing multiple open-source tools.

In this work, we present a toolbox for phage analysis (PhaBOX), a comprehensive web service for phage identification, lifestyle prediction, taxonomic classification, and host prediction. In PhaBOX, we optimize the integrated tools to improve the running speed and efficiency. In addition, PhaBOX can visualize the essential features that are important for making the final predictions, such as the similarity-based relationships between the input sequence and other phages, predicted proteins on the input sequences, and protein homology. PhaBOX can take either metagenomic assemblies or whole genome sequencing assemblies as inputs. All the predictions and intermediate results are provided for downstream analysis.

1.1 Overview

PhaBOX is a web server developed with Python/R, providing integrative identification and characterization for phage contigs in metagenomic data. Our algorithms behind PhaBOX were peer-reviewed and published, including PhaMer (Shang *et al.* 2022) for phage identification, PhaTYP (Shang *et al.* 2023) for lifestyle prediction, PhaGCN (Shang *et al.* 2021) for taxonomic classification, and CHERRY (Shang and Sun 2022) for host prediction. All these tools combined the strength of alignment-based strategies and deep learning models to learn different sequence-based features, including protein organizations, sequence homology, and protein–protein associations. Our methods outperform the available programs in each task based on our rigorous tests on highly diverged phages, short contigs, mock metagenomic data, and real metagenomic data. For example, according to a third-party view (Tang *et al.* 2022), the earlier version of CHERRY, named HostG (Shang and Sun 2021) has the best performance on predicting the phage–host

Received: April 25, 2023. **Revised:** July 5, 2023. **Editorial Decision:** July 27, 2023. **Accepted:** August 1, 2023

relationship on the genus level. In CHERRY, we not only further improved the accuracy but also supported host prediction at the species level.

PhaBOX provides integrated phage identification and analysis on one website. It has a modular design and thus supports two modes of running. The default mode allows users to conduct end-to-end analysis including all the supported functions. The other mode allows users to only run the modules they need, e.g. predicting the hosts for recently sequenced phages. As shown in Fig. 1, the input of PhaBOX is the FASTA file containing assembled contigs from metagenomic data or whole genome sequencing data. Then, PhaMer is applied for phage contig identification. By default, the identified phage contigs will be used as inputs for all modules including lifestyle prediction, taxonomic classification, and host prediction. For each module, the homology search is carried out against the NCBI RefSeq database. We updated the database and optimized the program to save computing resources. PhaBOX has a running time of approximately 50 s per 8 kb contig under its default mode. It is ∼70% of the total time of running each program individually. We also provided a multi-threads local version of PhaBOX for users who want to analyze phages in a large amount of metagenomic data.

In addition to achieving faster end-to-end phage analysis, PhaBOX provides detailed visualizations of the essential features behind the methodology, which is not available in the previous works. Figures 2 and 3 demonstrate the visualization of essential components, such as similarity-based relationships between contigs and other phages, predicted proteins on the contigs, and protein homology, which provide evidence to support the generation of predictions.

In the following sections, we will showcase how to use PhaBOX for identifying and characterizing phage contigs in metagenomic data. We will demonstrate the utility of PhaBOX and detail the visualization functions provided on the result page.

2 Case study

In this case study, we demonstrate that PhaBOX can simplify and speed up phage analysis in human microbiome data. The microbiome data are sequenced from fecal samples of 145 Chinese individuals, including 71 Type 2 diabetes (T2D) patients and 74 controls (Qin *et al.* 2012). We downloaded the assembled contigs provided by the authors and uploaded the FASTA files to the PhaBOX server. To maintain high-quality predictions, we set the minimum length of contigs as 10 kb as suggested in Qin *et al.* (2012). In total, we obtained 129 138 contigs in 145 samples. We used PhaBOX to identify and characterize the phages from the gut metagenomic samples.

2.1 Example of the result page

One example result page is shown in Fig. 2. There are a total of six modules on the result page, including the tab for downloading results, pie graphs of the prediction, contig results, phage family classification results, phage host prediction results, and low confidence results. The meaning of each module is listed below.

2.1.1 The module for downloading results

As shown in the top-right of Fig. 2, this module contains all the predictions and intermediate files for users to download. The standard outputs of PhaBOX are four CSV files containing the results of identified phages, their taxonomy, lifestyles, and hosts. PhaBOX also provides intermediate files of the essential features, including the homology of phage-related proteins output by PhaMer/PhaTYP and knowledge network output by PhaGCN/CHERRY. The homology files containing phage-related proteins in FASTA format and BLAST alignment results in CSV format. The knowledge network contains several types of interactions between metagenomic assemblies and phages/bacteria in the database, such as gene-sharing information, CRIPSRs, and sequence similarity. The network files are in standard input formats for most network-based visualization tools, such as Gephi (Bastian *et al.* 2009) and Cytoscape (Shannon *et al.* 2003).

2.1.2 Graphs of prediction

In the PhaBOX web server, pie graphs are employed to visualize the final prediction results and stacked bar graphs are provided as alternative options. There are eight charts in total,

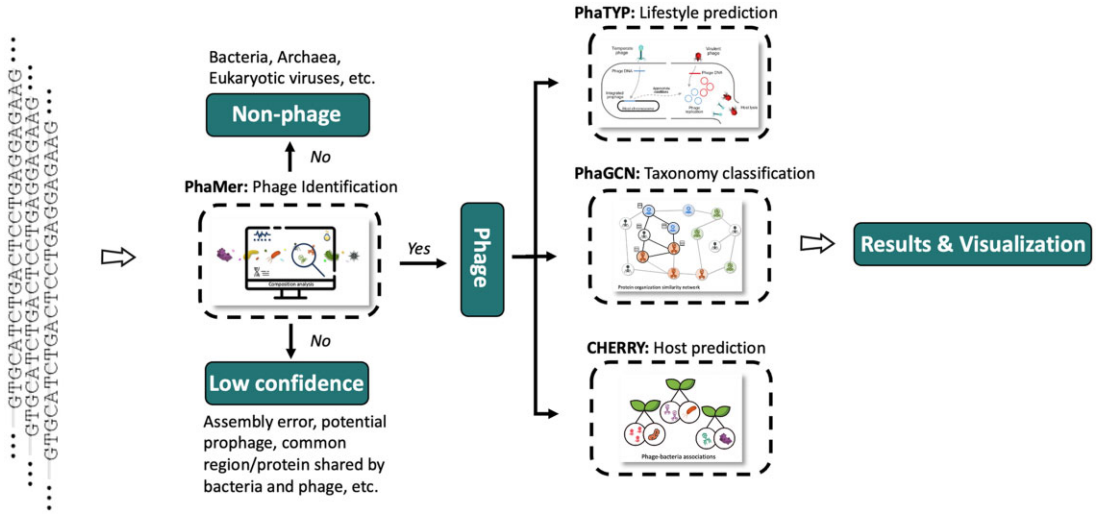

Figure 1. The pipelines of PhaBOX. The input of PhaBOX is the FASTA file containing assembled contigs. Then, PhaMer (Shang *et al.* 2022) is applied for phage identification. Only the contigs predicted as phages will be used for lifestyle prediction (PhaTYP; Shang *et al.* 2023), taxonomic classification (PhaGCN; Shang *et al.* 2021), and host prediction (CHERRY; Shang and Sun 2022). Finally, we provided the predictions and visualization for users on the result page.

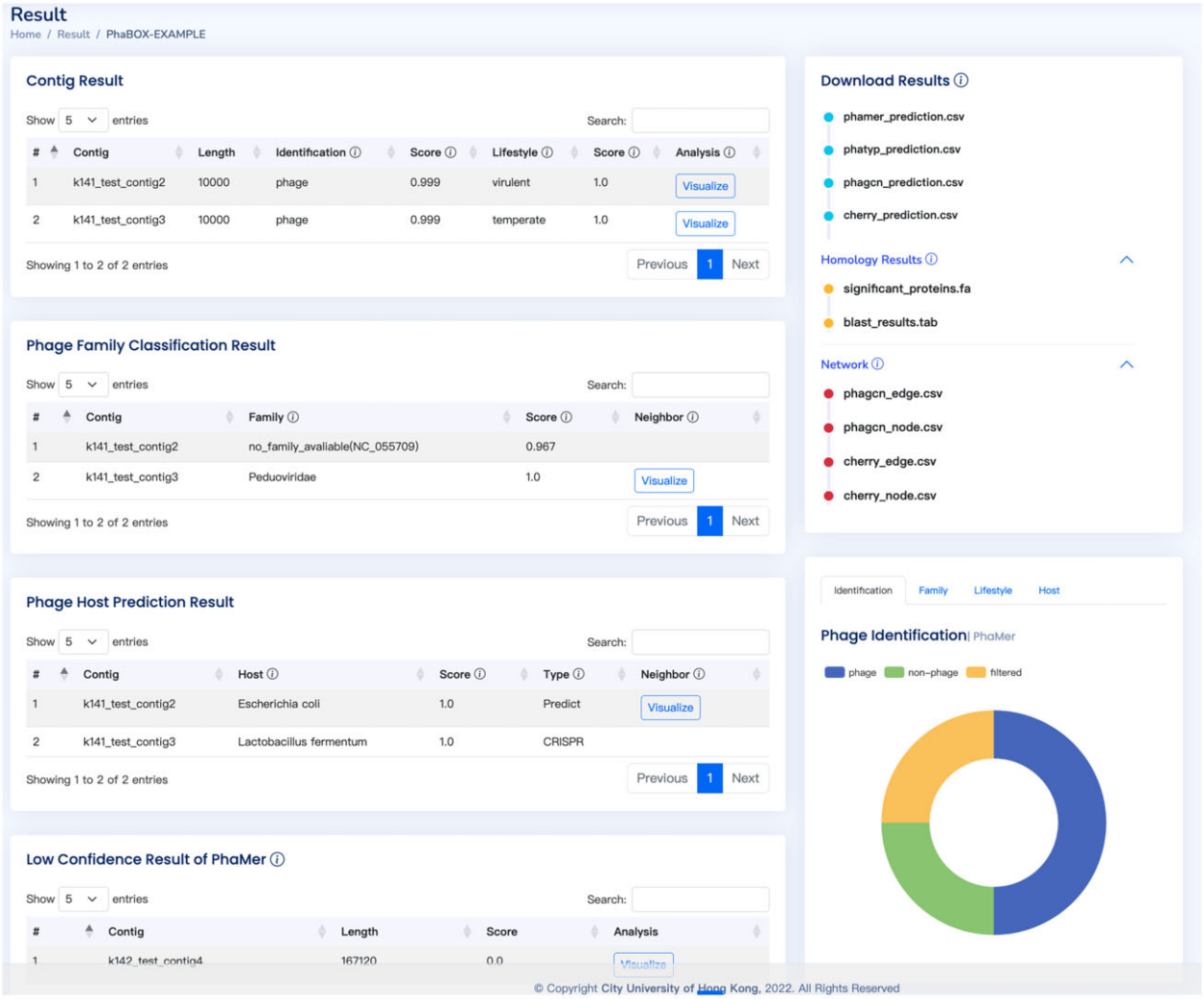

Figure 2. An example of the result page. The visualization can be accessed by clicking the visualize button. All the results can be downloaded from the top-right download panel. The meaning of the head will be shown when moving the cursor onto the icon of the header.

including the results of the percentage of identified phages, taxonomic composition, lifestyle composition, and host prediction. As shown in Fig. 3E and F, the pie graph and stacked bar graph show the percentage of identified phages.

2.1.3 Contig results

As shown in Fig. 2, the top-left table contains all the identified phages and their lifestyles predicted by PhaMer and PhaTYP. The visualization of the homology search on the input contigs can be accessed by clicking the "Visualize" buttons. Then, Fig. 3A will be presented to show the position of the translated proteins with their alignments score. Finally, each protein's detailed alignment results can be accessed as shown in Fig. 3B.

2.1.4 Phage family classification results and host prediction results

As shown in Fig. 2, the middle-left tables contain the family-level taxonomic classification results predicted by PhaGCN and host prediction results predicted by CHERRY. Because both methods utilize sequence similarity and gene organization as features to construct phage–phage and phage–bacterium relationship networks, we provide the visualization of these networks. However, the complete similarity network is too large to be visualized on a web page. Thus, we only show the one-step neighborhood of each contig as shown in Fig. 3C. The visualized network is interactive, and users can drag the nodes to check their accessions. As mentioned in *download results* module, the complete graph can be conveniently generated using our provided file and any of the commonly used network visualization tools such as Gephi. An example of a visualized complete graph by Gephi is shown in Fig. 3D.

2.1.5 Low confidence results

As shown in Fig. 2, the bottom-left table contains all contigs that have alignments with the database, but the confidence scores given by the model are too low to be predicted as phages. There are two possible reasons for the low confidence. First, there may exist assembly errors in the contigs. Second, the contigs contain regions that are shared by phage and bacteria. Thus, we provide the corresponding results and visualization for the user who wants to further analyze these sequences.

2.2 Outputs of PhaBOX

We uploaded all the T2D contigs to PhaBOX and analyzed the phage contigs. In addition, we recorded the running time of using PhaBOX and the total time of running each program separately. All the methods are run on Intel® Xeon® Gold 6258R CPU with 36 threads. The total time for PhaBOX to

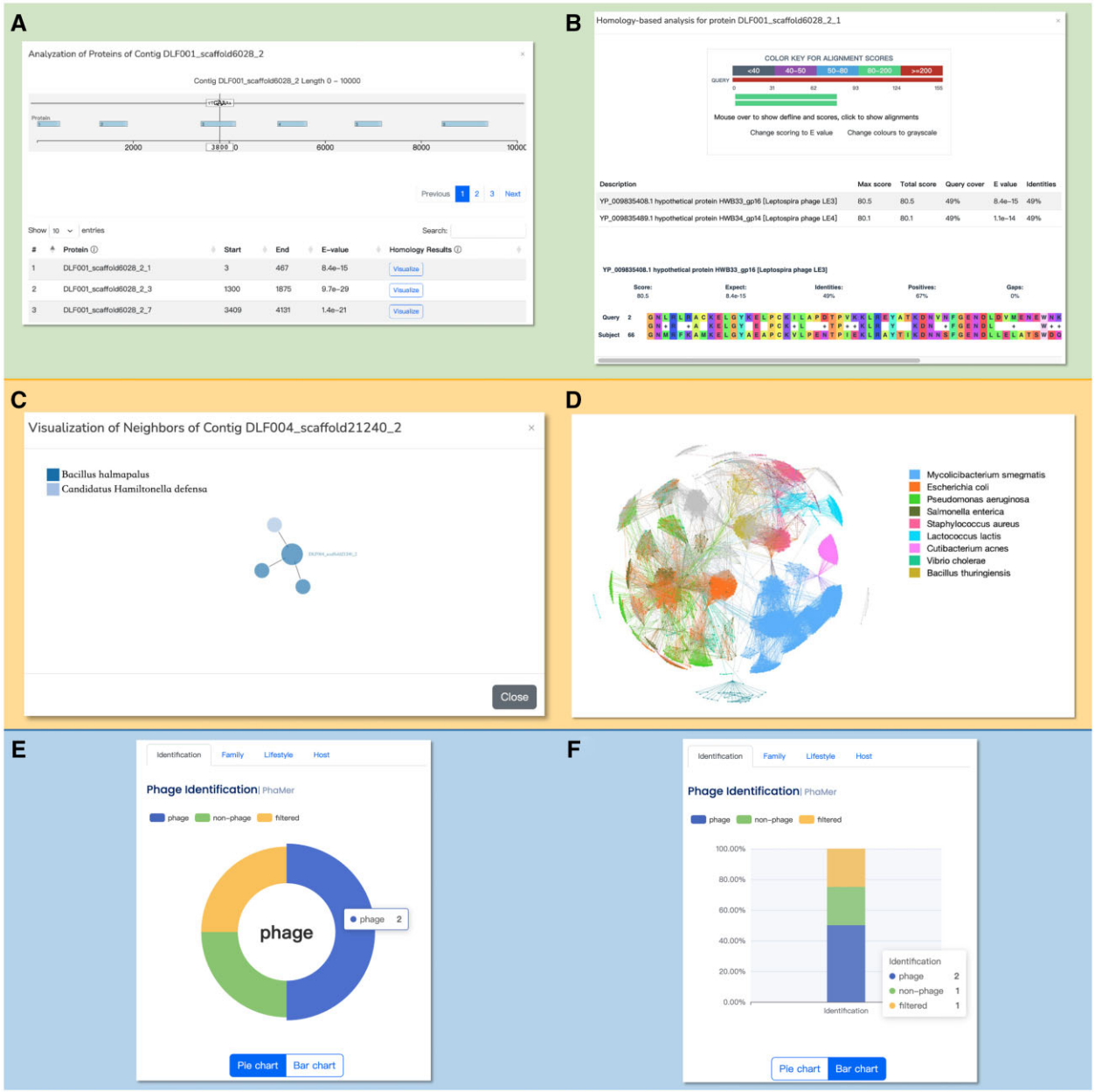

Figure 3. The visualization of the results. (A) The visualization of the homology search on the input contigs. The position of the translated proteins with high alignments score will be shown, and the detailed alignment results (B) can be accessed by clicking the visualize button. (C) We will show the one-step neighbors on the sequence-similarity graph to reveal the relationships between the input contigs and the phages/bacteria in the database. The complete sequence-similarity graph is provided in the "Network" panel in Fig. 2. They are in the standard format for most network-based visualization tools. One example using Gephi to visualize the graph is shown (D). (E) The pie graph of the prediction result. (F) The stacked bar graph is provided as an alternative option.

analyze 129 138 contigs in 145 samples is 216 min, saving 40% of the running time (328 min) of running each program separately.

Then, we summarized the prediction for each sample. There are 4851 phage contigs identified from the 145 samples. They belong to 18 families and infect 211 kinds of host species. PhaBOX provides multiple files for different functions. The summary file contains the comprehensive analysis for each contig (*prediction_summary.csv*). Table 1 shows an example of the summarized prediction. More detailed information about the prediction file, such as the confidence of the prediction, can be found in the supplementary files.

3 Methods

The architecture of the PhaBOX server consists of two major components: a client web interface and a server backend. The client web interface is responsible for submitting the tasks and displaying the output. It was implemented by JS, CSS, jQuery, Bootstrap, and their extension packages. Specifically, the sequence similarity was visualized by BlasterJS, the protein sequence viewer was presented using pViz, and the topological graph structure was drawn using Plotly in R. The server backend is responsible for interacting with users through the web interface, handling users' input, and executing the whole prediction process. The prediction pipeline

Table 1. Example summary file output by PhaBOX.[a]

Name	Length	Lifestyle	Taxonomy	Host
DLF001_scaffold16878_1	12 659	Virulent	Drexlerviridae	*Bacteroides fragilis*
DLF003_scaffold56673_2	51 717	Temperate	Casjensviridae	*Bacillus halmapalus*
DLM016_scaffold64767_1	15 261	Virulent	Salasmaviridae	*Lactobacillus plantarum*
NLM021_scaffold58236_5	38 364	Temperate	Straboviridae	*Lactobacillus gasseri*
NLF002_scaffold2199_6	28 963	Temperate	Straboviridae	*Clostridioides difficile*
NLF005_scaffold21492_3	14 868	Temperate	Ackermannviridae	*Streptococcus mutans*

[a] The results are generated using the four CSV files: phamer_prediction.csv, phagcn_prediction.csv, phatyp_prediction.csv, and cherry_prediction.csv.

contains four functional modules including phage identification, lifestyle prediction, taxonomic classification, and host prediction. In the phage identification and lifestyle prediction tasks, we adopt the state-of-the-art language model, Transformer (Vaswani *et al.* 2017), to automatically learn abstract patterns from the "language" of phages. In the taxonomic classification and host prediction tasks, we construct a knowledge graph by integrating multiple protein and DNA-based sequence features. Then the graph convolutional neural network is applied to utilize features from both labeled and unlabeled samples. All of the methods are comprehensively benchmarked with state-of-the-art tools on multiple datasets, including RefSeq dataset, low similarity dataset, metagenomic dataset, and etc. The former interface was implemented by the fast and lightweight Python-based Flask framework and the extension Python packages. The server backend employs a lite SQL database that stores and updates the job information and status. The scheduling method also allows the architecture to be added to add new computational facilities to meet the increasing demand for predicting ever-accumulating genome-scale data. More detailed information can be found under the "Home" tab on the PhaBOX web page.

4 Conclusion

Because phages ubiquitously exist in many different ecosystems, such as soil and marine samples, we develop this web server to accommodate the needs of users from different fields. It is expected that an easy-to-use web server can help advance the field of phage discovery in different types of ecosystems. The integration of phage identification, lifestyle prediction, taxonomic classification, and host prediction of our platform provides not only comprehensive analysis for metagenomic assemblies but also detailed visualizations for users. The case study demonstrates that PhaBOX can facilitate users with fast and convenient phage characterization in metagenomic data. Future improvements include adding more functions, such as novel protein annotation and protein structural analysis. We will also upgrade the hardware of our server to provide a faster prediction.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding

This work was supported by the City University of Hong Kong [Project 9667256 and 9678241]; the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

Data availability

All data and codes used for this study are available online via: https://phage.ee.cityu.edu.hk/download.

References

Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *Proc Int AAAI Conf Web Soc Media* 2009;**3**:361–2.

Cobián Güemes AG, Youle M, Cantú VA *et al.* Viruses as winners in the game of life. *Annu Rev Virol* 2016;**3**:197–214.

Cristobal-Cueto P, García-Quintanilla A, Esteban J *et al.* Phages in food industry biocontrol and bioremediation. *Antibiotics* 2021;**10**:786.

Fernández L, Rodríguez A, García P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J* 2018;**12**:1171–9.

Mushegian A. Are there 1031 virus particles on earth, or more, or fewer? *J Bacteriol* 2020;**202**:e00052-20.

Petrovic Fabijan A, Lin RC, Ho J *et al.* Safety of bacteriophage therapy in severe *Staphylococcus aureus* infection. *Nat Microbiol* 2020;**5**:465–72.

Qin J, Li Y, Cai Z *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60.

Roth KDR, Wenzel EV, Ruschig M *et al.* Developing recombinant antibodies by phage display against infectious diseases and toxins for diagnostics and therapy. *Front Cell Infect Microbiol* 2021;**11**:697876.

Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* 2021;**37**:i25–33.

Shang J, Sun Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol* 2021;**19**:1–15.

Shang J, Sun Y. CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model. *Brief Bioinform* 2022;**23**:bbac182.

Shang J, Tang X, Guo R *et al.* Accurate identification of bacteriophages from metagenomic data using transformer. *Brief Bioinform* 2022;**23**:bbac258.

Shang J, Tang X, Sun Y. PhaTYP: predicting the lifestyle for bacteriophages using BERT. *Brief Bioinform* 2023;**24**:bbac487.

Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.

Tang T, Hou S, Fuhrman JA *et al.* Phage–bacterial contig association prediction with a convolutional neural network. *Bioinformatics* 2022;**38**:i45–52.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: Guyon I, von Luxburg U, Bengio S (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017, 5998–6008.