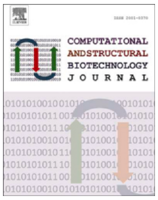
Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj



Research Article

GSPHI: A novel deep learning model for predicting phage-host interactions via multiple biological information

Jie Pan [a], Wencai You [a], Xiaoliang Lu [a], Shiwei Wang [a], Zhuhong You [b,*], Yanmei Sun [a,*]

[a] *Key Laboratory of Resources Biology and Biotechnology in Western China, Ministry of Education, Provincial Key Laboratory of Biotechnology of Shaanxi Province, The College of Life Sciences, Northwest University, Xi'an 710069, China*
[b] *School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China*

ARTICLE INFO

Article history:
Received 23 February 2023
Received in revised form 14 June 2023
Accepted 15 June 2023
Available online 16 June 2023

Keywords:
Phage-host interactions
Graph embedding technique
Deep neural network

ABSTRACT

Emerging evidence suggests that due to the misuse of antibiotics, bacteriophage (phage) therapy has been recognized as one of the most promising strategies for treating human diseases infected by antibiotic-resistant bacteria. Identification of phage-host interactions (PHIs) can help to explore the mechanisms of bacterial response to phages and provide new insights into effective therapeutic approaches. Compared to conventional wet-lab experiments, computational models for predicting PHIs can not only save time and cost, but also be more efficient and economical. In this study, we developed a deep learning predictive framework called GSPHI to identify potential phage and target bacterium pairs through DNA and protein sequence information. More specifically, GSPHI first initialized the node representations of phages and target bacterial hosts via a natural language processing algorithm. Then a graph embedding algorithm structural deep network embedding (SDNE) was utilized to extract local and global information from the interaction network, and finally, a deep neural network (DNN) was applied to accurately detect the interactions between phages and their bacterial hosts. In the drug-resistant bacteria dataset ESKAPE, GSPHI achieved a prediction accuracy of 86.65 % and AUC of 0.9208 under the 5-fold cross-validation technique, significantly better than other methods. In addition, case studies in Gram-positive and negative bacterial species demonstrated that GSPHI is competent in detecting potential Phage-host interactions. Taken together, these results indicate that GSPHI can provide reasonable candidate sensitive bacteria to phages for biological experiments. The webserver of the GSPHI predictor is freely available at http://120.77.11.78/GSPHI/.

1. Introduction

Available reports have suggested that bacterial infections may be rooted in the development and progression of various diseases [1], such as pneumonia [2], osteomyelitis [3], meningitis [4], endocarditis [5], and different types of cancers [6]. In addition, phage have been and continue to be the key to molecular biology since its origin [7]. The advent of antibiotics has allowed millions of patients with bacterial infections to be cured. However, due to the misuse of antibiotics, bacterial resistance continues to increase, which makes it more difficult to treat bacterial infections with antibiotics [8]. Moreover, this situation is exacerbated by the over-consumption and uncontrolled use of antibiotics. As the same time, bacteria can rapidly develop resistance to new antibiotics, which significantly reduces the effectiveness of antimicrobial drugs [9]. On the other hand, the high cost and long experimental circle make it difficult for companies to develop new antibiotics [10]. Therefore, development of novel approaches to promote resistance reversal and treat bacterial diseases is urgently needed [11].

Bacteriophages are the virus that specialises in infecting and killing bacterial cells [12]. Phages are reproduced by replication and proliferation, which produces many progeny phages and leads to bacterial cell lysis [13]. Additionally, phages can also replicate exponentially [14]. This summary of properties makes phages to be one of the most promising therapies for tackling the crisis of antibiotic resistance [15]. Identification of PHIs (phage-host interactions) serves to investigate whether phages can be used to treat bacterial infectious diseases or their symptoms [16]. However, traditional PHIs development experiments are often time-consuming, costly, and risky. Thus, some investigators attempt to develop computational

[*] Corresponding authors.
E-mail addresses: zhuhongyou@gxas.cn (Z. You), sunyanmei@nwu.edu.cn (Y. Sun).

https://doi.org/10.1016/j.csbj.2023.06.014

methods to predict PHIs, which can help to screen for target phages for synergistic treatment [17].

Ecological co-evolutionary processes produced phages and bacterial genomes and left their evolutionary information in the genomic sequences, so various computational approaches based on sequence information have been developed. For example, Ahlgren et al. [18] provided a computational tool called VirHostMatcher (VHM), which predicts PHIs by conducting a comprehensive evaluation of eleven oligonucleotide distance measures over various k-mers lengths. However, it is difficult to apply VHM to large data sets as it is too time-consuming. To deal with this problem, Galiez et al. [19] developed WIsH, which predicts the prokaryotic host based on the genome sequence of phages. Compared with VHM, WIsH models can reduce running times through the constructed Markov models. In addition, a number of machine learning based models have been adopted to identify PHIs [20], such as random forest (RF), logistic regression (LR), support vector machine (SVM), and naive Bayesian (NB). Furthermore, with the invention of PHP [21] and VIDHOP [22] model, the PHIs prediction accuracy has been brought to a new level. The PHP model calculated the differences between viruses and target hosts in the K-mers frequencies for the purpose to generated a Gaussian model. Recently, Several studies demonstrated that the receptor-binding proteins in the bacterial cell walls or membranes have a pivotal role in the phage adsorption processes [23,24]. Thus, some researchers started to focus on predicting PHIs based on receptor protein sequences. For instance, Li et al. [25] produced a novel deep learning-based approach to predict the hosts of phages from their sequence data. Leite et al. [26] proposed a method that used the primary sequences of phages and host proteins. Specifically, they combined sequence information with some machine learning classifiers (i.e RF, SVM, NB, and LR) to improve the prediction outcomes for phage-host interactions.

Despite these encouraging results, there are still some challenges. First, a massive number of experimentally identified PHIs pairs are collected in public domain databases [27], but only a few pairs are non-redundant that can be used to train the prediction models. This limitation makes it difficult to develop high-performance prediction tools. Secondly, most existing approaches only used the DNA sequence of phages or the protein sequence of hosts, and few studies have combined these impact factors together [28]. Thirdly, most of these machine learning models lack sufficient explanations for the mechanism of PHIs. Graph embedding algorithms have recently attracted growing attention in cell biology, and bioinformatics. Some attempts have been made to utilize such techniques to tackle different tasks. As a typical model for graph neural networks, Structural Deep Network Embedding (SDNE) applied the deep learning algorithms to learn network topological features. For example, Yi et al.: [29] collected nine biological macromolecule associations between diseases, miRNAs, drugs, proteins, and lncRNA to build a complex network. Then they used the SDNE algorithm to fully consider the topological information of these nodes. Meanwhile, some researchers developed deep neural network (DNN) to increase the interpretability of the predictive model. Advancements in these techniques have allowed us to provide further improvements of prediction accuracy.

In this work, we propose a novel PHIs prediction framework named GSPHI, based on a graph embedding technique SDNE and sequence fusion feature to deal with the questions of PHIs prediction. To be specific, we first constructed a phage-host interactions graph to summarize the connections between phages and bacteria. Nodes in the graph represent the phages and target hosts, and the links indicate their interactions. Then a graph neural network, SDNE, is used to capture the behavior features from their interaction links. Meanwhile, the natural language processing algorithm, word2vec, is applied to encode the tail protein and DNA sequence of phages and receptor-binding proteins of hosts to extract the attribute information. Finally, GSPHI integrate the behavior information with attribute information as a fusion matrix, and then implemented the prediction task using a deep neural network (DNN). Comparison results with the state-of-art machine learning classifiers and graph embedding methods demonstrated the high efficiency of our model. Case studies on three highly virulent pathogenic bacteria further justify the usefulness of the proposed model. The results of these comprehensive computational experiments indicate that GSPHI is very suitable for predicting phage-host interactions.

2. Methods and materials

2.1. Construction of a PHIs database focused on clinically relevant pathogens

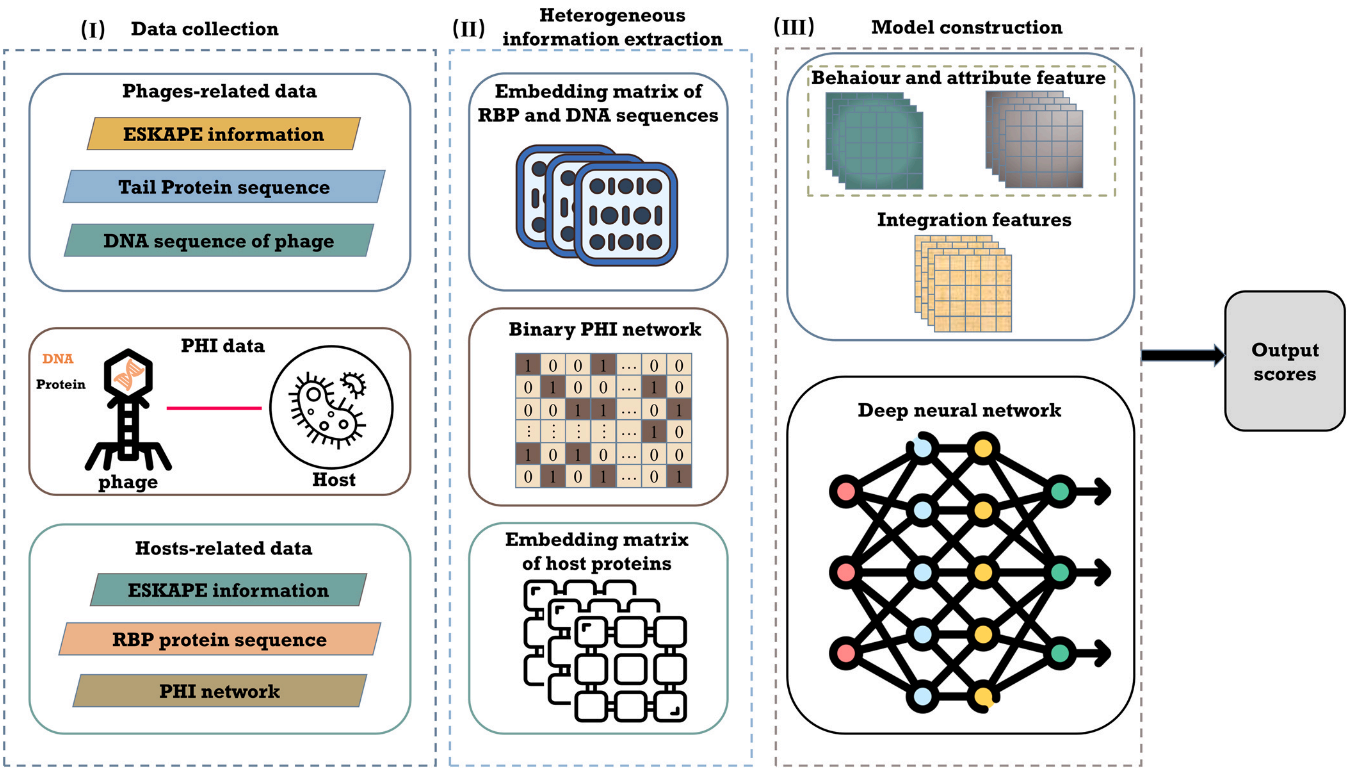
The tail protein of phages and receptor-binding protein (RBP) on the host surface determined whether the phage can adsorb on the host. Meanwhile, an essential function of phages DNA is to instruct the synthesis of their endogenous counterparts (tail protein). Hence, we took these three factors into account when we constructed the prediction model. According to this principle, we collected 1170 RBP sequences with the DNA and proteins information related to the tail structure of the target phages from three different public databases, including UniProtKB [30], UniRef [31] and MillardLab (http://millardlab.org). This dataset is dominated by ESKAPE (*Enterococcus faecium*, *Acineto-bacter baumannii*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, and *Enterobacter species*) pathogens, and supplemented with *Escherichia coli*, *Salmonella enterica* and *Clostridium difficile*. In clinics, these pathogenic bacteria are all highly invasive. The frequent misuse and overuse of antibiotics caused the lack of specific bacterial targeting ability for these bacteria. Identical sequences were removed in order to reduce computational load. Through this way, we finally collected 1232 phage-host interaction pairs consisting of nine bacterial species.

2.2. Overview of GSPHI

In this work, we proposed a deep learning framework named GSPHI, which can improve the performance of PHIs prediction by complementing multiple information. For a multi-perspective analysis, we combined the DNA and protein sequence information of the phage tail with the RBP sequence information of hosts. The whole model is illustrated in Fig. 1. Specifically, we first collected PHIs datasets about the lethal and highly drug-resistant bacteria. Then we used a novel graph embedding technique, SDNE, to extract behavior information from phage and host associations links (DNA-RBP and tail protein-RBP). The deep learning-based SDNE algorithm can extract both local and global information from heterogeneous PHIs networks. Meanwhile, we also applied the natural language processing technique, word2vec, to encode the sequence of DNA, and proteins of phage tail, and RBP for extracting information on attributes. In this way, abundant relationship of phages and hosts in kinds of bio-scale can be extracted. It is noteworthy that we developed a novel prediction model of GSPHI and constructed a heterogeneous interaction network to mine multimedia genetic relationships, which generates the topology-preserving representation of phages and hosts. In addition, large numbers of ablation experiments were performed to demonstrate the feasibility, acceptability, and effectiveness of the developed GSPHI.

2.3. Representing phage-host pairs with structural deep network embedding

In recent years, graph embedding techniques have received increasing attention in bioinformatics [32–34]. Compared with the conventionally analysis method, which is based on density



1	0	0	1	…	0	0
0	1	0	0	…	1	0
0	0	1	1	…	0	1
⋮	⋮	⋮	⋮	…	1	0
1	0	1	0	…	0	0
0	1	0	1	…	0	1

Fig. 1. The overview of GSPHI. (I) shows the PHIs dataset collecting. (II) shows the extraction of multiple information from PHIs network. (III) denotes the training module, and the output scores represent the potential PHIs pairs.

calculation, degree statistic, and multivariate clustering approaches, the graph embedding algorithms can represent the comprehensive features of the networks by generating a low-dimensional vector. Some researchers have reported that the structural deep network embedding (SDNE) algorithm delivers a competitive high prediction performance [35]. Therefore, in this work, we applied the SDNE algorithm to extract behaviour features of phage and hosts, and collected network proximity from each node. Different from shallow neural networks, SDNE utilized the core of DNN to model the complex nonlinear relationship between node representations. Since SDNE can retain both first and second-order proximity, it also has ability to preserve both local features and global representations from the PHIs network simultaneously. The whole algorithm can be separated into two segments: the first part is to model the first-order proximity, which was constructed by the supervised Laplace matrix. The second part is to model the second-order proximity through an unsupervised deep autoencoder. Finally, the middle layer outputs of the deep autoencoder were used as the representation of the nodes in the network, which maps the gene expression signatures to a low-rank feature space. The flow diagram of the SDNE algorithm is illustrated in Fig. 2. SDNE is composed of an encoder-decoder architecture, which has the same basic components as the deep autoencoder. The encoder part used R hidden layers to perform the non-linear transformations, which will map a low-dimensional representation x_i to a low-rank space, and $y_i{\cdot}x_i = G_i$ is used to describe the low-rank features. While the decoder cell is adopted to reconstruct the nodes representation, then the final vector is expressed as $\hat{x}_i$.

When SDNE applies the supervisory information to analyse the local structure of the interaction network, which was derived from the first-order proximity and the loss function is calculated by:

$$L_{1st} = \sum\nolimits_{i,j=1}^{n} G_{i,j} y_i^{(R)} - y_{j2}^{(R)2} = \sum\nolimits_{i,j=1}^{n} G_{i,j} y_i - y_{j2}^{2} \tag{1}$$

where the low-rank vectors of node i and node j can be described as y_i and y_j, respectively. The objection function of Eq. (1) borrows the ideal of Laplacian Eigenmaps [36], the penalties will be incurred when similar vertices are mapped far apart in the embedding space.

The second-order proximity is the ability to identify how similar neighbour-hood structure of a pair of vertexes are, and the unsupervised components develop to preserve the global network structures, and its objection function is denoted by:

$$L_{2nd} = \sum\nolimits_{i=1}^{n} (\hat{x}_i - x) \odot b_{i2}^{2} = (\hat{X} - X) \odot B_F^2 \tag{2}$$

where B represents a $N \times N$ matrix and $\odot$ denotes the Hadamard product between $(\hat{x}_i - x)$ and b_i. If $b_{i,j} = 1$, else $b_{i,j} = \beta > 1$, where β is free parameter and $\beta > 1$. $X = [x_1, x_2, ..., x_N]^T$, $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_N]^T$. In addition, in order to avoid overfitting, $L2$ -regularization term was established to optimate the parameters, and it should be calculated by:

$$L_{reg} = \frac{1}{2} \sum\nolimits_{R=1}^{R} \left(U_F^{2(r)} + \hat{U}_F^{(r)2} \right) \tag{3}$$

Here, R represents the total number of hidden layers, U^r and $\hat{U}^{(r)}$ denotes the weight matrices in the kth-layer. Finally, the SDNE algorithm integrates formula (1)-(3), and it can be shown as:

$$L_{mix} = L_{2nd} + \alpha L_{1st} + \nu L_{reg} = \left(\hat{X} - X \right) \odot B_F^2$$

$$+\alpha \sum\nolimits_{j,j=1}^{N} G_{ij} (y_i - y_j)_2^2 + \nu \frac{1}{2} \sum\nolimits_{R=1}^{R} \left(U_K^{2(r)} + \hat{U}_F^{2(r)} \right) \tag{4}$$

In this work, the SDNE algorithm was applied in the form of an adjacency matrix G to extract the behaviour information form the phage-host interaction network. Then we obtained a $(p + h) \times \tau$ embedding matrix MD, and MD_i is the row of MD, p and h correspond

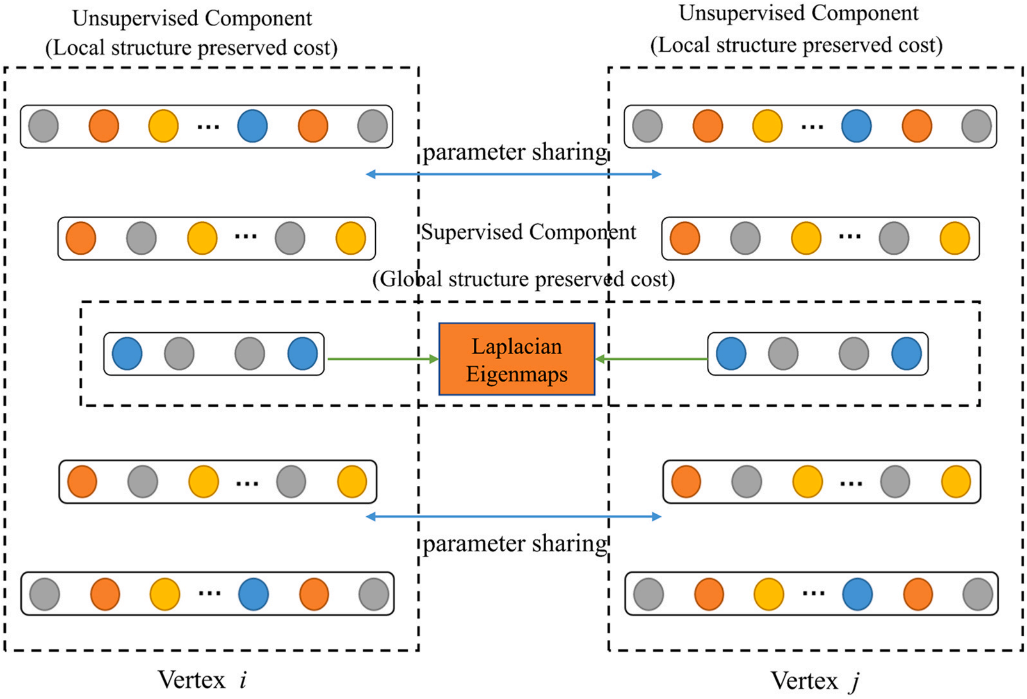


Fig. 2. The framework of SDNE algorithm.

to the property embeddings of phages and hosts, respectively. Meanwhile, τ is a free parameter that represents the embedding dimension of each node.

2.4. Embedding the biomolecular sequence by Word2Vec

In models of natural language processing (NLP), the goal of word embedding techniques, such as Word2vec [37] and doc2ver [38] is to learn a projection that represents words and documents as high-dimensional features and maps them into a rich lower-dimensional space. In the Word2vec algorithm, there are two approaches for representing the context of words: the Continuous Skip-Gram Model (Skip-Gram) and the Continuous Bag-of-Words Model (CBOW). The Skip-Gram model can identify the context based on the current word, while the CBOW-based method predicts the current word based on the context. When the sample size is not very large, Skip-Gram is more efficient and accurate than CBOW based method. However, CBOW learns words faster and more frequently. In this work., considering the size of the sequence series in this work, we adopted the CBOW model to represent the attribute features of PHIs nodes.

The amino acid sequence of phages and hosts was encoded into a matrix by Word2vec to abstract attribute characteristics from the PHIs network. The k-mers (k consecutive amino acids) approach was used to regard DNA, proteins of the tail, and RBP sequences as a single word and each sequence were expressed by numerous k-mers. For instance, given an RBP sequence MSTITQFPS, the units of 4-mers are MSTI, STIT, TITQ, ITQF, TQFP and QFPS. To increase computational speed, we used the python package genism [39] to train a CBOW-based model for the purpose of learning the appearance pattern of k-mers. In this work, the biological sequences and k-mers refer to the sentences and word, respectively. DNA and protein sequences of phages and hosts were encoded as embedding vectors in 64 dimensions. Since previous research [40] had proved that 4-mer provided the largest AUC by 5-fold cross-validation, we set k to 4. The details of the Word2vec algorithm used in RBP sequences are shown in Fig. 3.

2.5. Deep neural network

Deep learning is one of the most active fields in machine learning. Artificial neural network (ANN) was inspired by neural networks in the brain and consists of multiple layers of interconnected computed units[41–43]. The ANN model was designed with three layers: input, hidden, and output layer. The number of hidden layers determines the depth of the neural network, and the width corresponds to the maximum number of neurons in the hidden layer. With the continued increase of computational power, it became possible to train networks with larger numbers of hidden layers. The ANN model consists of many layers structure (two or more hidden layers) is called deep neural network (DNN). The function of DNN is not only to learn the high-level features from the original data, but also good at describing the complex structure of the high dimensional data.

In terms of its structure, DNN appears as a multilayer stack of plain modules. The features are first received in an input layer and non-linear transformations are then performed between the multiple hidden layers. The mean gradient will be calculated to adjust the corresponding design weights before producing the final outputs. Additionally, all neurons of the first hidden layer were connected to all neurons from the input layer, and all neurons of the last hidden layer were connected to the output layer. Then, a weighted sum of its input will be counted by the neurons and nonlinear activation functions are used to evaluate its output. In this work, rectified linear unit (ReLU) [44], tanh, and softmax [45] are employed as the activation functions. More specifically, the tanh function was used in input layer, while the activation function in the hidden layer and output layer were ReLU and softmax function, respectively. The binary cross entropy function was used as a loss function. Concurrently, the dropout learning algorithm [46] and Adam optimizer [47] were also employed to avoid overfitting and accelerate training. The entire network is defined as follows:

$$H_{i1}^{m} = \sigma_1(W_{i1}X_{i1} + b_{i1}),\ i = 1, ., n \tag{5}$$

$$H_{ij}^{m} = \sigma_1(W_{ij}H_{i(j-1)} + b_{ij}),\ j = 1, ..., n;\ j = 2, ..., h_1;\ m = 1, 2 \tag{6}$$

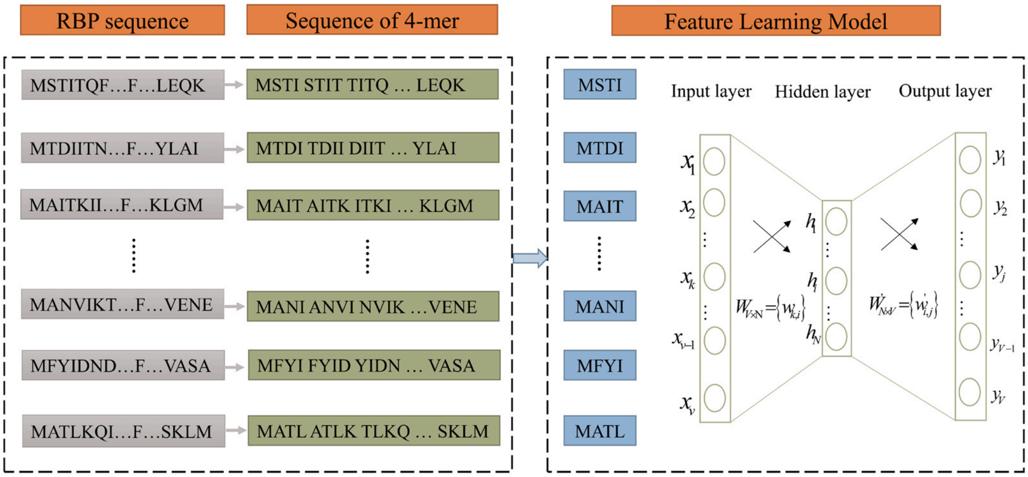


Fig. 3. Pipeline of the Word2vec model in 4-mer case.

$$H^3_{ik} = \sigma_1\left(W_{ik}\left(H^1_{ih_1} \oplus H^2_{ih_1}\right) + b_{ik}\right), i = 1, \ldots,n;\ \ k = h_1 + 1 \tag{7}$$

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln\left(\begin{array}{l}\sigma_2\left(W_{ih_2}H_{ih_2} + b_{ih_2}\right)+\\ (1-y_i)\ln\left(1-\sigma_2\left(W_{ih_2}H_{ih_2} + b_{ih_2}\right)\right)\end{array}\right)\right] \tag{8}$$

where m expresses the individual network, and n represent the batch sizes of PHIs pairs that are used for network training. The depth of the fused network and two individual networks is represented by h_2 and h_1. σ_1 and σ_2 represents the activation function ReLU and softmax for the hidden layer and output layer, respectively. X and H is the batch training input and output of corresponding layers. Variable W represents the weight matrices between the input layer, hidden layer, and output layer, and *b* is the bias term. In addition, the symbol ⊕is the concatenation operator, and y denotes the corresponding desired outputs.

2.6. Evaluation criteria

In this study, the 5-fold cross-validation (5-fold CV) was performed to computationally measure the prediction performance of GSPHI. We first randomly divided the ESKAPE dataset into five subsets with equal sample sizes, and then four of them were adopted as the training set and the remaining one as the test set. The process is repeated 5 times until each subset has been used as a test set once and only once. Finally, the average and standard deviations of these results were taken as the prediction results of GSPHI. In the experiment, accuracy (ACC.), sensitivity (Sen.), specificity (Spec.), precision (Prec.), and F1-score (F1) were adopted as assessment criteria for the predictive power of the GSPHI model. The details of their formulas are provided below:

$$ACC. = \frac{TP + TN}{FP + TP + FN + TN} \tag{9}$$

$$Sen. = \frac{TP}{FN + TP} \tag{10}$$

$$Spec. = \frac{TN}{TN + FP} \tag{11}$$

$$Prec. = \frac{TP}{TP + FP} \tag{12}$$

$$F1 = \frac{2 \times Prec. \times Sen.}{Prec. + Sen.} \tag{13}$$

where TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively. Meanwhile, we also plot the receiver operating characteristic (ROC) [48] and calculate the area under ROC curves (AUC value) [49] and PR (precision-recall) curves to demonstrate the performance of GSPHI.

3. Prediction result

3.1. Detailed descriptions of GSPHI model evaluation and performance

In the specific experiment, GSPHI employed the 5-fold-CV technique to generate the evaluation indicators on the ESKAPE data set, and the detailed experiment results are summarized in Table 1. We can see that the prosed model achieved an average prediction accuracy of 86.65 %, and its standard deviation was 1.55 %. On the evaluation indicators of Sen., Spec., Prec., F1-score and AUC, GSPHI achieved results of 88.40 %, 84.91 %, 85.43 %, 86.88 % and 0.9208, and their standard deviations were 1.81 %, 1.96 %, 1.74 %, 1.53 % and 0.0119, respectively. Fig. 4 presents the ROC and PR curves of the 5-fold CV generated by GSPHI on the ESKAPE data set.

3.2. Influence of different information on model performance

To assess the predictive ability and robustness of the GSPHI model, we performed experiments that only used the behaviour or attribute information. To be specific, we carried out ablation studies

Table 1
Results of 5-fold CV performed by GSPHI on ESKAPE dataset.

5-fold	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1 (%)	AUC
Fold-1	87.83	90.47	85.19	85.93	88.14	0.9255
Fold-2	85.70	86.82	84.58	84.92	85.86	0.9130
Fold-3	85.09	86.21	83.98	84.33	85.26	0.9173
Fold-4	85.90	89.05	82.76	83.78	86.33	0.9095
Fold-5	88.74	89.45	88.03	88.20	88.82	0.9393
Average	86.65 ± 1.55	88.40 ± 1.81	84.91 ± 1.96	85.43 ± 1.74	86.88 ± 1.53	0.9208 ± 0.0119

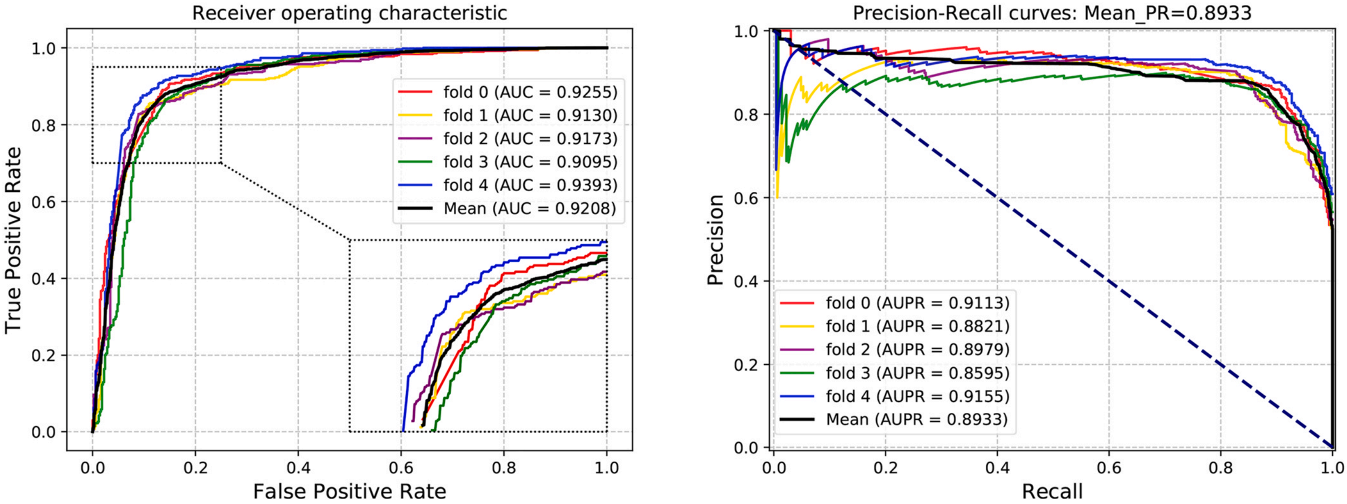


Fig. 4. ROC and PR curves of 5-fold CV achieved by GSPHI on ESKAPE data set.

to investigate the contribution of these components. Hence, we designed the following eight variants of GSPHI, which help us to analyse the proposed approach from multiple perspectives.

- DR-A is a variant that only used the attribute information of the DNA-RBP sequence.
- TR-A is a variant that only used the attribute information of the Tail-RBP sequence.
- DTR-A is a variant that combined the attribute information of the DR-A and TR-A.
- DR-B is a variant that only used the behaviour information of the DNA-RBP sequence.
- TR-B is a variant that only used the behaviour information of the Tail-RBP sequence.
- DTR-B is a variant that combined the behaviour information of the DR-B and TR-B.
- DR-AB is a variant that combined information of the DR-A and DR-B.
- TR-AB is a variant that combined the information of the TR-A and TR-B.

Table 2 shows the result of the proposed model and its eight variants, and the detailed 5-fold CV results of these eight variants are summarized in Supplementary Table S1–S8. We found that the performance was lower when any module was removed, which implies that all the components are essential for the proposed model. The performance of TR-B shows the largest observed, with the ACC and AUC decreasing by 28.75% and 0.2693, respectively. The comparison results between DR-AB and DR-A showed that the accession of behaviour information can effectively improve prediction performance. Furthermore, the results of comparative analysis for the DR-AB and TR-AB variant showed that the features derived from DNA sequences were more useful than those derived from tail protein sequences. This scenario may attribute to the fact that the phage DNA directs the synthesis of its tail proteins.

3.3. Comparison of different machine learning-based classifiers

The deep neural network is a multi-layer feedforward neural network, which was used in the GRASDENPHI model to accurately estimate the phage-host interaction. In our experiments, to independently validate the influence of a neural network-based classifier module, we compare it with some different machine-learning based classifiers. In particular, the fusion features were generated by the behaviour, and attribute information was remained unchanged, and only replace the DNN module with some widely-used machine learning classifiers, including support vector machine (SVM), K-nearest neighbour (KNN), random forest (RF) and Gradient Boosting Decision Tree (GBDT) algorithms to build some conventional models to predict host-interacted phages. The results of these predictive performances were summarized in Table 3. For a more visual comparison of these experimental results, we used box plots (Fig. 5) to describe the superiority of the proposed model. As is observed from Fig. 5, GSPHI has yielded excellent results on ACC, Sen, F1, and AUC values. From these computational results, the proposed model represented the strongest competitiveness. The experimental result demonstrated that the DNN module adopted in GSPHI has a powerful tendency to determine whether phages have interacted with bacterial, which is favourable to improving overall predictive performance. The detailed 5-fold CV results for these machine learning classifiers were reported in Supplementary Material (Tables of S9–S12).

Table 2
Results of 5-fold CV performed by GSPHI on ESKAPE dataset.

Method	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1 (%)	AUC
DR-A	80.48 ± 3.52	87.20 ± 3.67	73.77 ± 3.83	76.89 ± 3.22	81.72 ± 3.30	0.8444 ± 0.0344
TR-A	76.11 ± 4.21	85.26 ± 8.83	66.96 ± 6.66	72.20 ± 3.43	77.98 ± 4.61	0.8212 ± 0.0328
DTR-A	83.39 ± 1.54	87.14 ± 1.50	79.63 ± 2.19	81.08 ± 1.75	83.99 ± 1.42	0.8908 ± 0.0173
DR-B	81.78 ± 3.56	88.58 ± 4.70	74.98 ± 2.57	77.94 ± 2.63	82.91 ± 3.52	0.8890 ± 0.0279
TR-B	57.81 ± 1.63	52.06 ± 12.73	63.56 ± 10.51	59.16 ± 2.05	54.47 ± 7.94	0.6335 ± 0.0126
DTR-B	75.05 ± 0.86	83.04 ± 0.51	67.06 ± 1.84	71.61 ± 1.10	76.90 ± 0.61	0.8151 ± 0.0052
DR-AB	82.02 ± 3.74	88.83 ± 3.99	75.22 ± 3.99	78.21 ± 3.41	83.17 ± 3.53	0.8771 ± 0.0332
TR-AB	81.82 ± 3.32	88.74 ± 3.14	74.90 ± 3.62	77.96 ± 3.07	83.01 ± 3.07	0.8937 ± 0.0273
GSPHI	86.65 ± 1.55	88.40 ± 1.81	84.91 ± 1.96	85.43 ± 1.74	86.88 ± 1.53	0.9208 ± 0.0119

Table 3
The performance of different classifiers under 5-fold CV performed on the ESKAPE dataset.

Method	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1 (%)	AUC
SVM	83.19 ± 0.75	78.13 ± 1.63	88.24 ± 1.13	86.94 ± 1.00	82.29 ± 0.90	0.8975 ± 0.0095
KNN	73.65 ± 0.83	81.49 ± 2.21	65.81 ± 1.97	70.46 ± 0.88	75.55 ± 0.95	0.8304 ± 0.0051
RF	82.30 ± 1.10	75.65 ± 1.91	88.95 ± 1.02	87.25 ± 1.08	81.03 ± 1.31	0.8698 ± 0.0194
GBDT	80.69 ± 2.21	80.27 ± 1.55	81.11 ± 3.36	81.00 ± 2.83	80.62 ± 2.01	0.8942 ± 0.0173
GSPHI	86.65 ± 1.55	88.40 ± 1.81	84.91 ± 1.96	85.43 ± 1.74	86.88 ± 1.53	0.9208 ± 0.0119

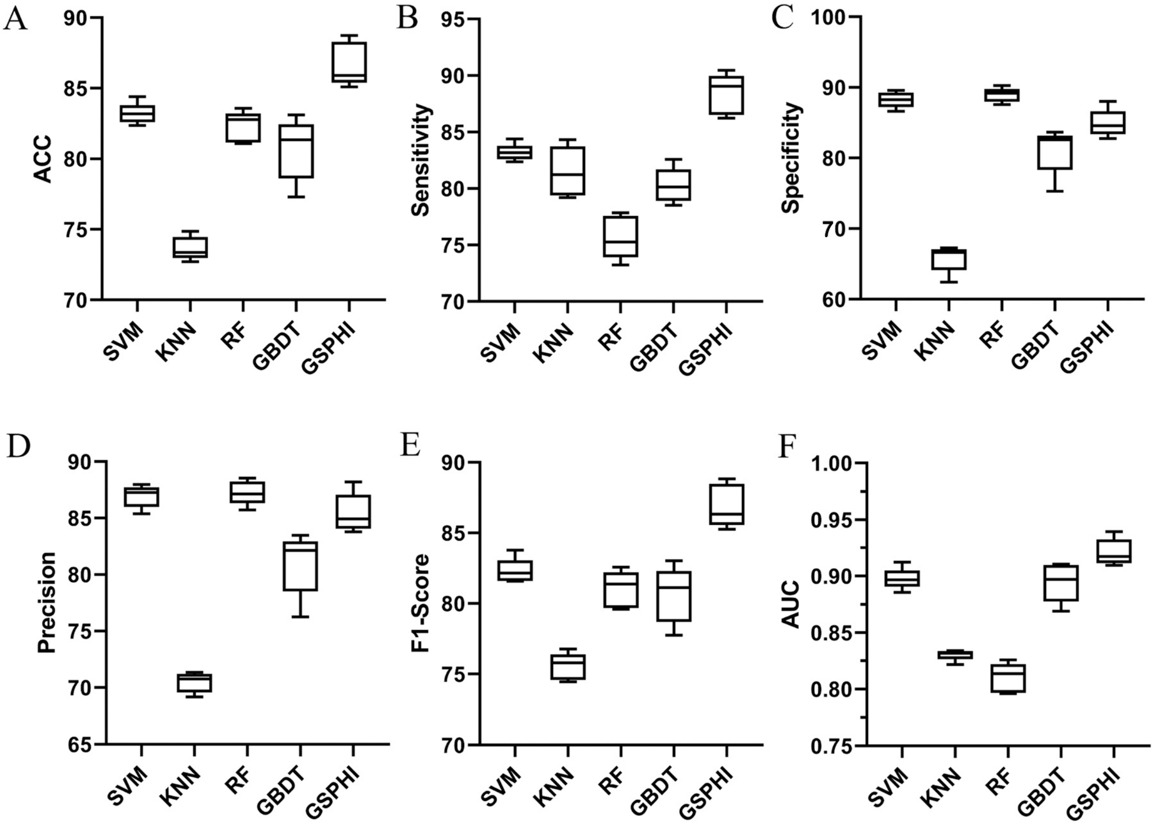


Fig. 5. Box plots of comparison results from different classifier. The X-axis represents the adopted machine-learning classifiers. The Y-axis represents the different evaluation metrics.

3.4. Comparison with some different graph embedding methods

To further demonstrate the prediction power of SDNE algorithm, various graph embedding methods were compared.

- Deepwalk [50]: Deepwalk is a deep learning method that vectorizes vertices in the graph and represents the potential relationship of the vertices.
- Line [51]: Line algorithm is a novel algorithm for large-scale network computing and it is suitable for any type of network structure.
- Hope [52]: Hope algorithm obtains the embedding vector by factorizing the similarity matrix between nodes.
- Laplacian Eigenmaps (Lap) [53]: Lap is a geometric algorithm, which was proposed to embed data samples from a low-dimensional manifold in higher dimensional space.

For a fair comparison, all methods are test with the same data dimension and the same DNN structure. Table 4 lists the prediction result of these four powerful graph algorithms. From Table 4, it was shown that the performance of all these methods with different graph algorithms are lower than our method. To visually compare the performance of these methods, we also plotted their ROC and PR curves in Fig. 6. For the performance of two factorization-based methods (Lap and Hope), as the scores of ACC, Spec, and Prec, obtained by Lap were better by 1.14 %, 5.76 %, and 2.63 % than those of Hope on ESKAPE dataset, respectively. However, despite these good prediction results performed by Lap algorithm, it was still 6.49 % lower than our method in terms of ACC values. The details results are reported in Supplementary Material (Tables 13–16), which suggested that our deep-learning based SDNE algorithm can improve the performance of the model in this experiment.

3.5. Case study: A. baumannii, P. aeruginosa, and S. aureus

With the intention of further assessing whether GSPHI could exhibit accurate and robust performance, we conducted three case studies with Gram-positive and Gram-negative bacteria, including *A. baumannii*, *P. aeruginosa*, and *S. aureus*. Where *A. baumannii* and *P. aeruginosa* belong to the genus of Gram-negative bacteria, and they usually cause pneumonia, meningitis, peritonitis, endocarditis, as well as urinary tract and skin infections. While *S. aureus* refers the Gram-positive bacterium, which is responsible for about 25 % of food poisoning. Specially, we used the data of all known phage-host interactions as training samples and the method of GSPHI to perform prediction. Then, we selected the top 20 phages in ascending order

Table 4
comparison results with some widely used graph embedding methods.

Method	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1 (%)	AUC
Deepwalk	79.03 ± 1.96	85.23 ± 2.74	72.82 ± 2.36	75.83 ± 1.81	80.24 ± 1.93	0.8633 ± 0.0188
Line	81.89 ± 0.66	87.43 ± 0.99	76.35 ± 2.16	78.74 ± 1.34	82.84 ± 0.41	0.8890 ± 0.0054
Hope	79.02 ± 1.12	89.57 ± 2.22	68.48 ± 3.77	74.04 ± 1.86	81.03 ± 0.76	0.8684 ± 0.0021
Lap	80.16 ± 1.84	86.08 ± 0.87	74.24 ± 3.56	77.03 ± 2.48	81.29 ± 1.47	0.8690 ± 0.0185
GSPHI	**86.65 ± 1.55**	**88.40 ± 1.81**	**84.91 ± 1.96**	**85.43 ± 1.74**	**86.88 ± 1.53**	**0.9208 ± 0.0119**

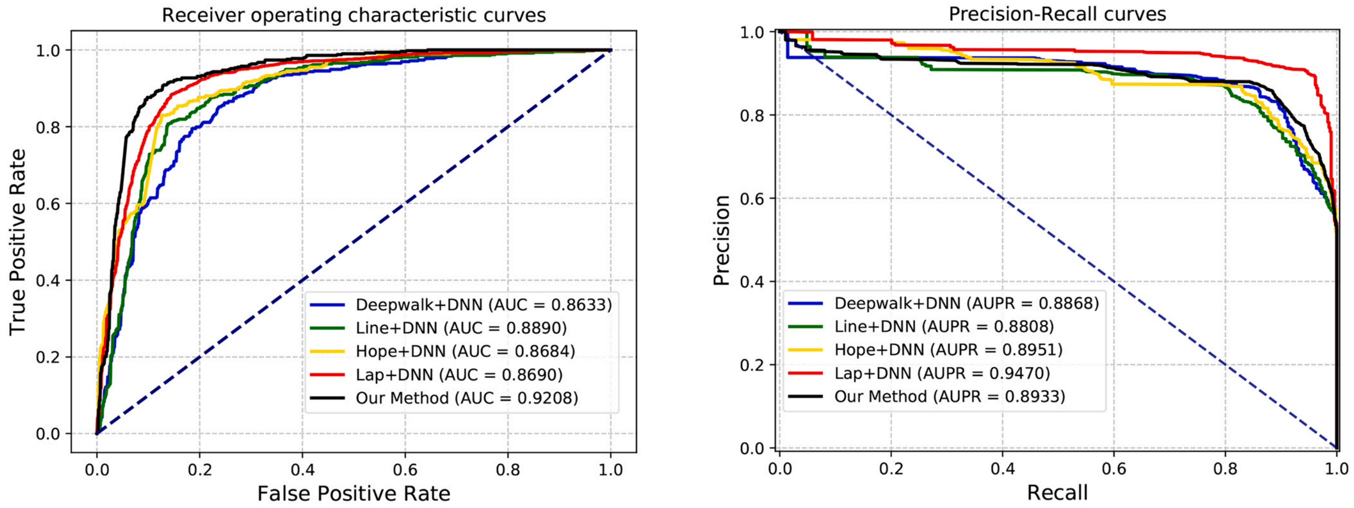


Fig. 6. Comparison results of different graph embedding methods about ROC and PR curves.

Table 5
The top 20 interactions of *Acinetobacter baumannii* related phages.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	JX976549	Confirmed	11	JX297445	N.A.
2	KU935715	Confirmed	12	EU734174	N.A.
3	KM672662	Confirmed	13	MF033348	Confirmed
4	KJ473423	Confirmed	14	KC862301	N.A.
5	KT804908	Confirmed	15	HQ186308	Confirmed
6	LN610572	Confirmed	16	MF001356	N.A.
7	LN881736	N.A.	17	KJ628499	Confirmed
8	EF151185	N.A.	18	JQ965645	N.A.
9	MH042230	Confirmed	19	KT454805	N.A.
10	MF033347	Confirmed	20	HE806280	Confirmed

Table 6
The top e20 interactions of *Pseudomonas aeruginosa* related phages.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	KT887559	Confirmed	11	KT184311	N.A.
2	MF158046	N.A.	12	HM035025	N.A.
3	MH688040	N.A.	13	KT372698	Confirmed
4	MH536736	Confirmed	14	KC862297	Confirmed
5	KU297675	Confirmed	15	KP994390	Confirmed
6	MK050846	N.A.	16	KX171209	Confirmed
7	FM887021	Confirmed	17	AJ505558	Confirmed
8	KF981730	N.A.	18	KM411959	Confirmed
9	KX171210	Confirmed	19	HG934469	N.A.
10	LN610575	Confirmed	20	AM265638	Confirmed

of probability for giving a detailed analysis. The detail results were shown in Tables 5–7. It was noted that 12, 13 and 11 out of the top 20 phages were verified in published database. It is noteworthy that other interactions that are not verified by the current literature. However, we could not rule out the possibility that there are some interactions between them. These case studies confirmed that GSPHI is an excellent model for examining implicit phage-host interactions.

Table 7
The top 20 interactions of *Staphylococcus aureus* related phages.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	KJ888149	Confirmed	11	KY794641	Confirmed
2	MH107769	Confirmed	12	DQ904452	N.A.
3	KY581279	Confirmed	13	HM137666	N.A.
4	KM606994	N.A.	14	JX080305	Confirmed
5	MH844529	Confirmed	15	KU867876	N.A.
6	MG656408	Confirmed	16	KF582788	N.A.
7	FR852584	Confirmed	17	AP011113	N.A.
8	AF406556	N.A.	18	KR902361	Confirmed
9	KR908644	Confirmed	19	KP687432	Confirmed
10	AF208841	N.A.	20	FJ839693	N.A.

4. Discussion and conclusion

Designing phage therapies or understanding the co-evolution of host-virus is important for addressing the continued emergence of antibiotic resistance. The identification of phage hosts is extremely critical for exploring whether phage can be employed to treat bacterial diseases. In this study, we proposed a deep learning framework of GSPHI to predict potential PHIs. This model can construct feature descriptors by fusing DNA and protein sequence information of the target phage tail with the RBP sequence information of hosts. In addition, to build an accurate prediction model, this model also considers the behavior information of the PHIs network to fuse multiple attributes into a deep neural network to effectively predicting potential phage-host pairs. Experimental results indicated that GSPHI is superior to other existing methods. Meanwhile, analyses of three pathogenic bacteria-associated phages case studies verify the possibility of application in the future.

The great predictive performance of GSPHI is mainly credited to three factors: (i) GSPHI integrates various biological information including natural language features and the behavior information from the PHIs network. (ii) Perform the deep learning algorithm SDNE to integrate the local and global topological information of the

PHIs network. (iii) GSPHI adopts DNN as the classifier to identify the interactions between phages and target hosts, which can efficiently identify the possible phage-interacted bacteria from multiple sequence information.

However, despite these impressive results, the proposed model has still faced some limitations. First, this model extracts the attribute and behavior features by natural language and deep learning algorithms. These features may not be interpretable and difficult for users to understand. Secondly, the samples that picked at random will bring some errors to the prediction results, as this may cause a little noise for these samples. In future work, we will attempt to incorporate sequence similarity information between phages and hosts to build a more accurate and robust model. Additionally, we would try to collect more biological attributes to construct a more comprehensive microbial information network that the proposed model can extract more expressive features of phages and bacteria. However, the complexity of the interaction network will increase with the increasing of biological data, and the redundancy and noise in functional information may become more severe, thus creating novel difficulties to the attribute fusion capabilities of the proposed model. To solve this problem, we would try to design an end-to-end network to provide comprehensively curated information in future work. As a complex network architecture that is designed to predict phage-host interactions, we hope that the proposed GSPHI model can provide new insights into the treatment of bacterial infectious diseases. The source code can be freely found from Github (https://github.com/NWUJiePan/Code).

Funding

This work was supported in part by the grants from the Science & Technology Fundamental Resources Investigation Program (Grant No. 2022FY101100) and National Natural Science Foundation of China (32170114 and 31770152). This work was also supported by the Science and Technology Innovation 2030—New Generation Artificial Intelligence Major Project (No. 2018AAA0100103).

CRediT authorship contribution statement

Jie Pan: Conceptualization, Methodology, Resources, Writing - review & editing. Wencai You: Data curation, Software. Xiaoliang Lu: Investigation, Project administration, Supervision, Validation. Shiwei Wang and Yanmei Sun: Formal analysis, Funding acquisition, Visualization. Zhuhong You: Funding acquisition, Roles/Writing - original draft. All authors have read and agreed to the published version of the manuscript.

Data Availability

GSPHI is also publicly available as an online predictor at: http://120.77.11.78/GSPHI/ All the code and dataset are available at https://github.com/NWUJiePan/Code.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.06.014.

References

[1] Wallis RS, O'Garra A, Sher A, Wack A. Host-directed immunotherapy of viral and bacterial infections: past, present and future. Nat Rev Immunol 2023;23(2):121–33.

[2] Meawed TE, Ahmed SM, Mowafy SM, Samir GM, Anis RH. Bacterial and fungal ventilator associated pneumonia in critically ill COVID-19 patients during the second wave. J Infect Public Health 2021;14(10):1375–80.

[3] Bottagisio M, Coman C, Lovati AB. Animal models of orthopaedic infections. A review of rabbit models used to induce long bone bacterial infections. J Med Microbiol 2019;68(4):506–37.

[4] Oordt-Speets AM, Bolijn R, van Hoorn RC, Bhavsar A, Kyaw MH. Global etiology of bacterial meningitis: a systematic review and meta-analysis. PloS One 2018;13(6):e0198772.

[5] Loyola-Rodriguez JP, Franco-Miranda A, Loyola-Leyva A, Perez-Elizalde B, Contreras-Palma G, Sanchez-Adame O. Prevention of infective endocarditis and bacterial resistance to antibiotics: a brief review. Spec Care Dent 2019;39(6):603–9.

[6] van Elsland D, Neefjes J. Bacterial infections and cancer. EMBO Rep 2018;19(11):e46632.

[7] de Jonge PA, Nobrega FL, Brouns SJ, Dutilh BE. Molecular and evolutionary determinants of bacteriophage host range. Trends Microbiol 2019;27(1):51–63.

[8] Andersson DI. Persistence of antibiotic resistant bacteria. Curr Opin Microbiol 2003;6(5):452–6.

[9] Andersson DI, et al. Antibiotic resistance: turning evolutionary principles into clinical reality. FEMS Microbiol Rev 2020;44(2):171–88.

[10] Towse A, Hoyle CK, Goodall J, Hirsch M, Mestre-Ferrandiz J, Rex JH. Time for a change in how new antibiotics are reimbursed: development of an insurance framework for funding new antibiotics based on a policy of risk mitigation. Health Policy 2017;121(10):1025–30.

[11] Maffei E, et al. Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection. PLoS Biol 2021;19(11):e3001424.

[12] Cisek AA, Dąbrowska I, Gregorczyk KP, Wyżewski Z. Phage therapy in bacterial infections treatment: one hundred years after the discovery of bacteriophages. Curr Microbiol 2017;74:277–83.

[13] Ongenae V, Mabrouk AS, Crooijmans M, Rozen D, Briegel A, Claessen D. Reversible bacteriophage resistance by shedding the bacterial cell wall. Open Biol 2022;12(6):210379.

[14] Li X-Y, Lachnit T, Fraune S, Bosch TC, Traulsen A, Sieber M. Temperate phages as self-replicating weapons in bacterial competition. J R Soc Interface 2017;14(137):20170563.

[15] Pires DP, Costa AR, Pinto G, Meneses L, Azeredo J. Current challenges and future opportunities of phage therapy. FEMS Microbiol Rev 2020;44(6):684–700.

[16] Núñez-Sánchez MA, et al. Characterizing phage-host interactions in a simplified human intestinal barrier model. Microorganisms 2020;8(9):1374.

[17] Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev 2016;40(2):258–72.

[18] Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res 2017;45(1):39–53.

[19] Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics 2017;33(19):3113–4.

[20] Zhang M, Yang L, Ren J, Ahlgren NA, Fuhrman JA, Sun F. Prediction of virus-host infectious association by supervised learning methods. BMC Bioinforma 2017;18(3):143–54.

[21] Lu C, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biol 2021;19(1):1–11.

[22] Mock F, Viehweger A, Barth E, Marz M. VIDHOP, viral host prediction with deep learning. Bioinformatics 2021;37(3):318–25.

[23] Alguwaizani S, Park B, Zhou X, Huang D-S, Han K. Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. J Healthc Eng 2018;2018.

[24] Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. PLoS Comput Biol 2020;16(5):e1007894.

[25] Li M, et al. A deep learning-based method for identification of bacteriophage-host interaction. IEEE/ACM Trans Comput Biol Bioinforma 2020;18(5):1801–10.

[26] Leite DMC, et al. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE,; 2018. p. 1818–25.

[27] Gao NL, et al. MVP: a microbe–phage interaction database. Nucleic Acids Res 2018;46(D1):D700–7.

[28] Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 2017;5:1–20.

[29] Yi H-C, You Z-H, Guo Z-H, Huang D-S, Chan KC. Learning representation of molecules in association network for predicting intermolecular associations. IEEE/ACM Trans Comput Biol Bioinform 2020.

[30] Bairoch A, et al. The universal protein resource (UniProt). Nucleic Acids Res 2005;33(suppl_1):D154–9.

[31] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23(10):1282–8.

[32] Guo L-X, et al. A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. Brief Bioinforma 2022;23(5):bbac391.

[33] Gong Y, Niu Y, Zhang W, Li X. A network embedding-based multiple information integration method for the MiRNA-disease association prediction. BMC Bioinforma 2019;20(1):1–13.

[34] Liu S, et al. Enhancing drug-drug interaction prediction using deep attention neural networks. IEEE/ACM Trans Comput Biol Bioinforma 2022.

[35] Wang D, Cui P, Zhu W. Structuraldeep network embedding. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 1225–34.

[36] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 2003;15(6):1373–96.

[37] Church KW. Word2Vec. Nat Lang Eng 2017;23(1):155–62.

[38] Le Q, Mikolov T. Distributed representations of sentences and documents. International Conference on Machine Learning. PMLR,; 2014. p. 1188–96.

[39] Rehurek R, Sojka P. Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer,; 2010.

[40] Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J 2020;18:153–61.

[41] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65(6):386.

[42] Farley B, Clark W d. Simulation of self-organizing systems by digital computer. Trans IRE Prof Group Inf Theory 1954;4(4):76–84.

[43] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5(4):115–33.

[44] Nair V, Hinton GE. Rectifiedlinear units improve restricted boltzmann machines. Icml. 2010.

[45] Li Z, Li H, Jiang X, Chen B, Zhang Y, Du G. Efficient FPGA implementation of softmax function for DNN applications. 2018 12th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE,; 2018. p. 212–6.

[46] Baldi P, Sadowski P. The dropout learning algorithm. Artif Intell 2014;210:78–122.

[47] Zhang Z. Improved adam optimizer for deep neural networks. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Ieee,; 2018. p. 1–2.

[48] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(4):561–77.

[49] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30(7):1145–59.

[50] Perozzi B, Al-Rfou R, Skiena S. Deepwalk:Online learning of social representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014. p. 701–10.

[51] Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. Line:Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web. 2015. p. 1067–77.

[52] Ou M, Cui P, Pei J, Zhang Z, Zhu W. Asymmetric transitivity preserving graph embedding. Proceedings of the 22nd ACM SIGKDD internationalconference on Knowledge discovery and data mining. 2016. p. 1105–14.

[53] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. Adv Neural Inf Process Syst 2001;14.