IGNITE: IGNITE: Identification of bacteriophaGes-prokaryotic associatioN UsIng Transformer with multi-layer pErceptron

Feng Zhou 24520220157289[1]; Muxi Li 24520220157279[1]; Jia Yi 24520220157285 [1]
Class XinYuan
Xiamen University
Xiamen, China

Abstract

Bacteriophages/Phages are viruses that infect bacteria and archaea and are key players in the microbial community. But in natural communities, the existence of a large number of phage-virus associations is not generally known. To understand their regulatory roles in various ecosystems and to harness the potential of bacteriophages for use in therapy, more knowledge of phage-host association is required. Therefore, characterizing phages-prokaryotic association is a critical component to understanding how biological systems work. Traditionally, virus research has used culture-based isolation techniques that provide direct identification of phage-host association. However, identifying the association between phage and prokaryotic based on traditional methods is time-consuming and labor-intensive. With the advent of metagenomic sequencing technologies, a large number of computational methods have been developed to infer the hosts of new viruses. Despite some promising results, computational host prediction remains a challenge because of the limited known interactions. In this work, we propose a method called IGNITE using Transformer and multi-layer perceptron (MLP) to predict phage-host association. First, we use a state-of-the-art natural language processing (NLP) model, the transformer, to extract features from phage. Then, IGNITE learns the feature representation of the host through a two-layer neural network. Finally, IGNITE uses a three-layer perceptron as a decoder to calculate the relative likelihood of phage and host.

Introduction

Viruses are the most abundant and highly diverse biological entities on Earth (Breitbart and Rohwer 2005)(Breitbart et al. 2002). Prokaryotic viruses which include phages and archaeal viruses play an important role in balancing the global ecosystem by regulating the composition of bacteria and archaea in water and soil (Ahlgren et al. 2017)(Lu et al. 2021)(Gregory et al. 2019). Viruses that infect bacteria referred as bacteriophages or phages in short have played essential rules in natural environments, they directly influence gut health and are associated with several human diseases, such as diabetes and Crohn's disease.

Researchers discovered antibiotics in 1928 and have since used them in clinical practice to treat serious bacterial diseases and save countless lives. However, because bacteria are highly adaptable, they can rapidly evolve resistance to new antibiotics, which will significantly reduce the effectiveness of the drug. Currently, phage therapy is a promising approach that uses viruses to infect and kill bacteria. Upon phages' recognition of specific types of receptors on the bacterial surface, they inject their DNA into the bacteria, resulting in replication, generation of additional phages, and production of an enzyme that dissolved the outer bacterial cell membrane to release the generated phages (Burstein et al. 2016)(Zhang et al. 2017)(Guerin et al. 2018). Therefore, a fundamental step in using phages to treat bacterial infection is to identify the hosts of phages, which will provide the key knowledge of using phages as potential antibiotics. Besides phage therapy, identifying the hosts of the novel phages have other applications such as gene transfer search, disease diagnosis, and novel bacterial detection. Currently, the method for determining the viral host is either to culture the virus which is low-throughput, time-consuming, and expensive, or to computationally predict the viral hosts which needs improvements at both accuracy and usability.

There are two major challenges for computational host prediction (Liu et al. 2019). The first one is the lack of known virus–host interactions. For example, the number of known interactions dated up to 2020 only accounted for ˜0.4 (1940) of the prokaryotic viruses at the NCBI RefSeq at that time. Meanwhile, among the 60 105 prokaryotic genomes at the NCBI RefSeq, only 223 of them have annotated interactions with the 1940 viruses. The limited known interactions require carefully designed models or algorithms for host prediction. Second, although sequence similarity between viruses and hosts is an insightful feature for host prediction, not all viruses share common regions with their host genomes. For example, in the RefSeq database, ˜0.24 viruses do not have significant alignments with their hosts. Therefore, a new prediction method is needed to calculate virus-host interactions.

In recent years, deep learning technology has received extensive attention in the field of bioinformatics, and researchers have applied such techniques to handle different tasks. Therefore, in this course assignment, our group plans to use deep learning-based methods to predict phgae-host association.

Related works

This section surveys the latest literature on predicting bacteriophage host association on basis of learning methods. Compared with alignment-based methods, learning-based methods are more flexible at predicting virus-host interactions for newly identified viruses. For example, several learning-based methods utilizes k-mer features for prediction. VirHostMatcher (VHM) utilized k-mer based oligonucleotide frequency (ONF) to predict the host of a selected virus by the greatest ONF similarity (Ahlgren et al. 2017). Galiez et al. then developed a homogenous Markov model called WIsH. They computed the likelihood of contigs and predicted de novo the host with the highest likelihood for virus host prediction and acquired higher accuracy and faster run-time compared with VHM (Galiez et al. 2017). PHP, a Gaussian model, was implemented using the differences of k-mer frequencies as features. It inferred the virus' label based on the highest probability from learned Gaussian distribution and gave a host prediction accuracy of 34% on VHM data set (Lu et al. 2021). VirHostMatcher-Net (VHM-net) was an advanced version of VirHostMatcher, which employed Markov random field framework while integrating CRISPR, alignment-free similarity measures. VHM-net got 59% and 86% accuracy at genus and phylum levels on 1462 known virus-host interaction pairs, respectively (Wang et al. 2020). Leite et al. utilized the primary protein structure sequences of bacteriophages and host and constructed several traditional machine learning models, including k-nearest neighbor (KNN), RF, SVM, and artificial neural network (ANN) to predict phage host association (Leite et al. 2018a)(Leite et al. 2018b). In Leite's method, negative pairs were randomly selected from the putative negative set, which may cause bias.

Latest learning-based method, mainly deep learning models, has achieved better performance and are capable of predicting association at finer classifications, such as species level prediction. For instance, vHULK models host prediction as a multi-class classification issue with prokaryotes serving as the labels and viruses serving as the inputs. It constructed a multi-layer perceptron model in which predicted protein sequences generated by genome sequences are searched against pVOGs database and help make predictions (Amgarten et al. 2020). Li et al. developed PredPHI for phage-host interaction prediction based on a deep convolution neural network. They collected protein sequence data and extracted features AAC and AC representing protein-related information to construct the model. Specially, PredPHI K-Means clustered the negative pairs to generate negative samples in training set and had 0.69 accuracy on test set.(Li et al. 2020) Similar to PredPHI, DeepHost also trained a CNN model for host prediction but based on genome. DeepHost encoded phage genomes into 3D matrices and applied spaced k-mers, which tolerates SNPs and InDels. DeepHost can achieve good performance, but only when predicting genomes within hits in BLAST.(Ruohan et al. 2022) Shang et al. presented HostG, a semi-supervised model to predict the hosts of prokaryotic viruses. They constructed a knowledge graph depicting virus-virus and virus-host similarity and applied graph convolution network for GCN learning. (Shang and Sun 2021)

Though learning methods have made great progress in predicting bacteriophage host association, there are still some remaining problems. First of all, although some BLAST-based methods can predict some reliable associations, as the data increases, the computational complexity will increase exponentially. Second, the lack of positive samples makes the training process difficult. Only a limited number of positive phage-host association pairs identified in the database were available for model training. The above studies most employ DNA sequences or protein sequences related features as model inputs. How to reasonably select the features of bacteriophages and hosts is another major issue.

Proposed Solution

In this paper, we regard the host prediction task as a link prediction task, where phage and host are encoded by different models, respectively. To be specific, For phages, we employ a state-of-the-art text embedding model, Transformer, to automatically learn feature representations from the "language" of phages. At the same time, the kmer method is used to obtain the embedding from the host, and then a two-layer multi-layer perceptron (MLP) is used to learn the feature representation of the host. Then the link prediction task can be defined as: given a phage embedding *pi* and a bacteria embedding *bi*, what is the probability of *pi* and *bi* having a link (infection). In the following section, We will first introduce how to translate the phage genome into protein sentences based on each phage. Then, the structure of IGNITE will be described in detail.A schematic diagram of IGNITE is shown in Figure 1.

Encoding into protein-based sentences

In the process of encoding phage, each token is derived from a protein cluster, which contains homologous protein sequences from genome phages. We construct protein clusters from genomic data from phage, where gene finding and protein translation are used on the downloaded DNA genomes. A recent study found that Prodigal is the best tool for identifying genes in viruses, particularly phages (González-Tortuero et al. 2021). Therefore, we used Prodigal to predict open reading frames in our training and test genome data using the default settings. Afterwards, we performed all-against-all DIAMOND BLASTP (Buchfink, Xie, and Huson 2015) on the predicted proteins and created a protein similarity network using protein pairs with an E-value of 1e-3 or lower. In this network, the proteins are represented as nodes and the alignments are represented as edges, with the edge weight indicating the E-value of the corresponding alignment. Finally, we used the Markov clustering algorithm (MKL) (Enright, Van Dongen, and Ouzounis 2002)with default parameters to group similar proteins into clusters, discarding any clusters containing fewer than two proteins. The process of constructing protein clusters is shown in Figure 2.

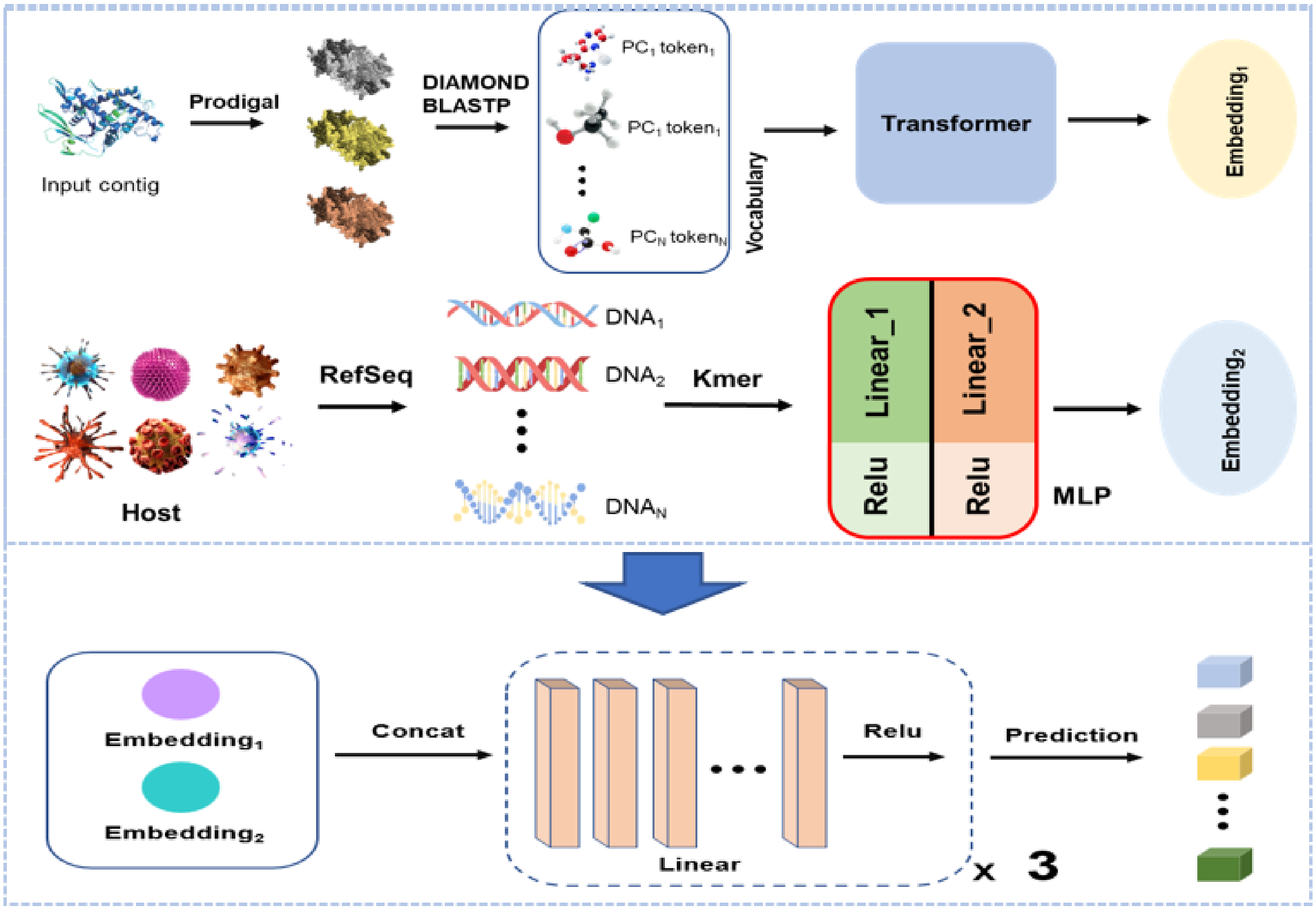

Figure 1: An overview of IGNITE.

We will use the generated protein clusters as tokens in our vocabulary to represent phages as sentences. We will also record the identification number of each token (protein cluster) and its position in the query sequence, as shown in Figure 1. Since the genome lengths of different phages are different, the resulting protein-based sentences are also of different lengths. We adhere to the guidelines outlined in the Transformer paper (Oliver et al. 2018) and limit the maximum length of a sentence to 300. If a sequence has more than 300 protein clusters, we only consider the first 300. For sequences with less than 300 tokens, we add zeros to the end of the sentence. Then, we create a 300-dimensional vector for the input sequence, where each dimension represents a unique token ID.

The Transformer model

Two main components in Transformer contribute to these aims: (1) the embedding layers and (2) the self-attention mechanism. As shown in Figure 3, the process of embedding the sentence and token positions for the Transformer block involves two layers: the protein-cluster embedding layer and the positional embedding layer. The protein-cluster embedding layer, similar to a lookup table, converts the input token into a numerical vector. However, due to the large size of the vocabulary (45,577), using one-hot encoding can result in sparse vectors, which can negatively impact the model's performance. To avoid this issue, we utilize a fully connected layer to conduct linear projection and generate a lower-dimensional embedding vector for each token. This FC layer acts as a learnable dictionary, mapping an ID of a token to its corresponding embedding vector. The model architecture of the transformer is shown in Figure 4.

Because Transformer contains no recurrence or convolution, it uses the positional embedding to encode the position information. In the model, we obtain the feature representation of the embedding layer as follows:

$$\begin{cases} I_s = FC(I_s, W_{Is}) \\ I_p = FC(I_p, W_{Ip}) \\ X = I_s + I_p \end{cases} \tag{1}$$

Which I_s is the input sentence and I_p is the position index vector for the input tokens. $W_{Is} \in R^{N*embed}$ and $W_{Ip} \in R^{len*embed}$ are the learnable parameters of the lookup table for protein-cluster embedding and positional embedding, respectively. N is the number of protein clusters, which is 45 577 in our model, and len is the maximum length

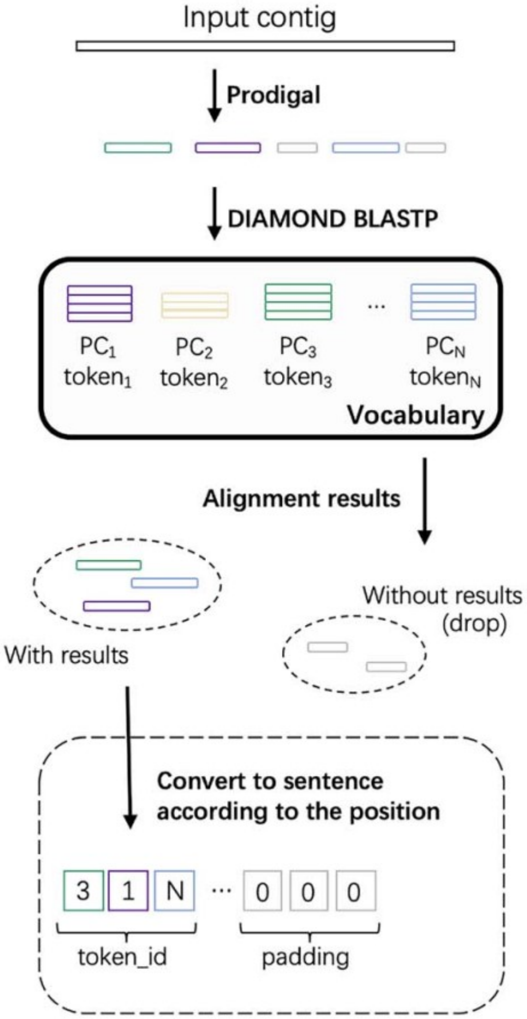

Figure 2: Converting inputs into protein-token sentences

of the sentence, which is 300 by default. embed is a hyperparamter of the embedding dimension and it is set to 512 by default following the guideline in (Oliver et al. 2018). Then, X will be fed into the Transformer block. Ideally, these embedding layers will capture some of the semantics of the input by placing semantically similar tokens close together in the embedding space.

After embedding the sentences, each token is converted into a vector of size 512 and the embedded sentences will be a $\mathbf{R}^{300*512}$ matrix. Then, we feed the matrix into the self-attention mechanism. We want to train a model to learn: given a set of proteins (query), which proteins (key) are usually co-present in phage genomes (value).

$$Attention(Q,K,v) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Eqn.2 show how the self-attention mechanism works. First, the embedded matrix X is projected by three FC layers into Q, K and V, respectively. Because the attention matrix only contains pairwise protein cluster information, to model different combinations of pairwise relationships, we use h FC layer groups for linear projections. Each group is called a head ($head_i$), and on each of these projected versions of queries Q_i, keys K_i and values V_i, we can perform the self-attention mechanism in parallel. To reduce computational complexity, in each FC layer we will reduce the dimension for the projected features. The dimension of the output will be $lend_s$, where ds is calculated by $embed/h$. In this work, we choose $h = 8$ by default. Thus, the formula of each head attention can be written as in Eqn.3.

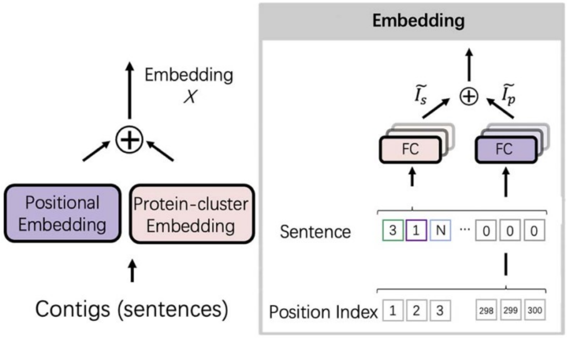

Figure 3: Converting inputs into protein-token sentences

$$\begin{cases} headi_i = Attention\left(Q_i, K_i, V_i\right) \\ Q_i = FC\left(X, W_i^Q\right) \\ K_i = FC\left(X, W_i^K\right) \\ V_i = FC\left(X, W_i^V\right) \end{cases} \tag{3}$$

The parameters in the FC layers are projections matrices: $W_i^Q \in R^{N\times ds}, W_i^K \in R^{N\times ds} and W_i^V \in R^{N\times ds}$Finally, we will concatenate the output from each head and form the final output of the multi-head attention block as shown in Eqn. 4, where $W^O \in R^{hd_s \times embed}$.

$$\begin{gathered} MultiHead(Q,K,V) = \\ FC\left(Concat\left(head_1,\ldots,head_h\right), W^0\right) \end{gathered} \tag{4}$$

Finally, we feed the output of the multi-head attention block to a two-layer neural network. The output of the neural network is the final feature representation for each phage.

Feature representation for prokaryotic

Because the genome sequence of bacteria is much larger than that of phage. Therefore, for the feature representation

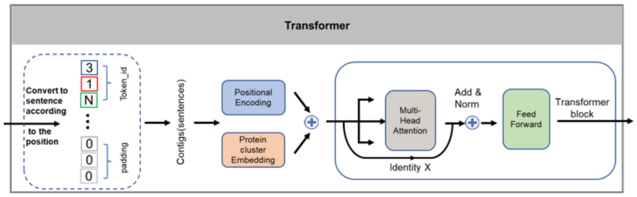

Figure 4: The model architecture of the transformer

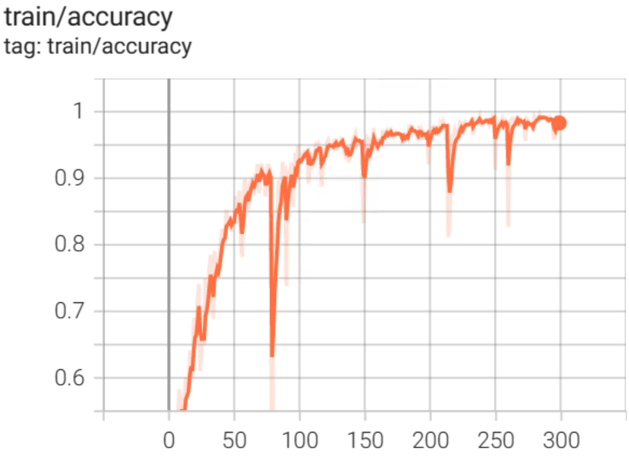

Figure 5: The classification accuracy of the training set

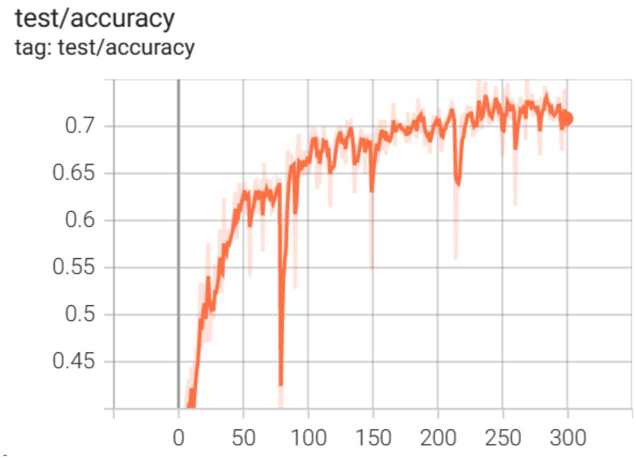

Figure 6: The classification accuracy of the test set

of prokaryotes, we first use kmer to obtain the initial features of prokaryotes, and then use two layers of fully connected layers to learn the initial features of hosts to obtain the final feature representation of each prokaryote.

Decoder for link prediction

After obtaining the characterization of the phage and host, we apply a three-layer neural network classifier to decode the embedded vectors outputted from the encoder. This decoder aims to judge how likely these query pairs form actual infections. Thus, the input of the decoder is a query set Q, and the output of the decoder is a probability score. Each element in Q is called a query vector q_ij and is calculated by Eq.5.

$$q_{ij} = encoder\left(p_i\right) - encoder\left(h_j\right) \tag{5}$$

First, we generate all-against-all virus–prokaryote pairs and calculate all query vectors $q_{ij} \in Q$. Then we employ a two-layer neural network to decode the feature vector for each input q_{ij} as shown in Eq.6.

$$\begin{cases} q_{ij}^{(l+1)} = \phi\left(q_{ij}^{(i)}\theta^{(l)}\right) \\ decoder\left(q_{ij}\right) = sigmoid\left(q_{ij}^{(L-1)}\right) \end{cases} \tag{6}$$

which decoder$(\cdot)$ represents the output of the link prediction decoder.Because the activation function of the output layer is the sigmoid function, decoder$(\cdot)$ can be used as the probability score for each pair.

Model training

Research shows that end-to-end learning can effectively improve the learning efficiency of the model. Therefore, our overall trainable parameters of IGNITE optimized by backpropagation loss. The trainable parameters of IGNITE are: (i) The weights of the Transformer and the two-layer perceptron during the encoding process and (ii) query parameter matrices in the decoder. There are two kinds of query pairs that will be generated by Eq. 5: positive pairs and negative pairs. Positive pairs represent known virus–prokaryote interactions given by the dataset. Negative pairs represent the pairs with no evidence for interaction. The training loss of the model is defined as follows:

$$J(i,j) = -\log P_r^{ij} - E_{n\sim p_r} \log\left(1 - P_y^{ik}\right) \tag{7}$$

During the training process, we optimize the model using the crossentropy loss as shown in Eq. 7. Because we form all-against-all query pairs from all viruses and hosts, the number of negative query pairs will be much larger than the positive query pairs. To solve this problem, rather than sampling a subset of the negative pairs, we optimize the model through negative sampling.

Experiment

In this experiment, we ulitized the benchmark dataset (the VHM dataset) introduced in (Lu et al. 2021) comprising 1940 viruses and 206 hosts. We download all 1940 viruses from the NCBI RefSeq database and separate the training set and test set according to their submission time (before and after 2015). Thus, we have 1306 positive pairs for training and 634 positive pairs for testing, respectively. Although every virus is unique, some of them infect the same host. During the training phase, each phage together with its known host, were treated as a positive phage-host pair, while the selection of a negative pair is achieved by combining this known host and random screening of one phage that does not actually interact with the host. Here, we trained our model IGNITE, evaluated our experimental results and compared our model against other 9 virus-host interaction prediction tools in terms of training accuracy and test accuracy over distinctive taxonomic levels (from species to phylum). During training process, we optimized several trainable parameters. Figure 5 and Figure 6 respectively show the classification accuracy of the training set and test set of the model at the Family level, which are recorded by the tensorboard module.

In experiment, we first trained our model and evaluated model performance over distinctive taxonomic levels from

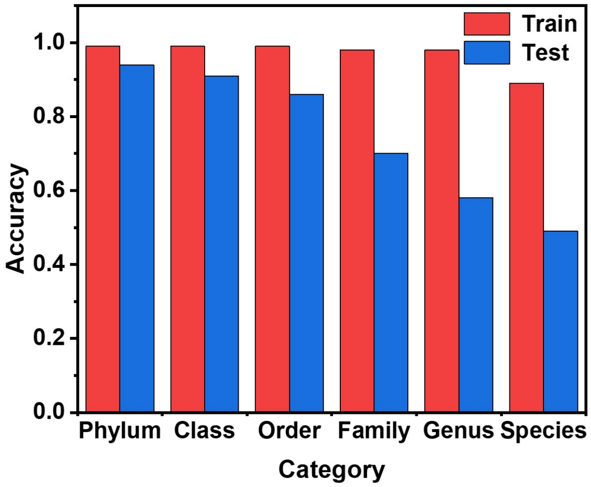

Figure 7: The comparison of the results of our method on the training set and test set

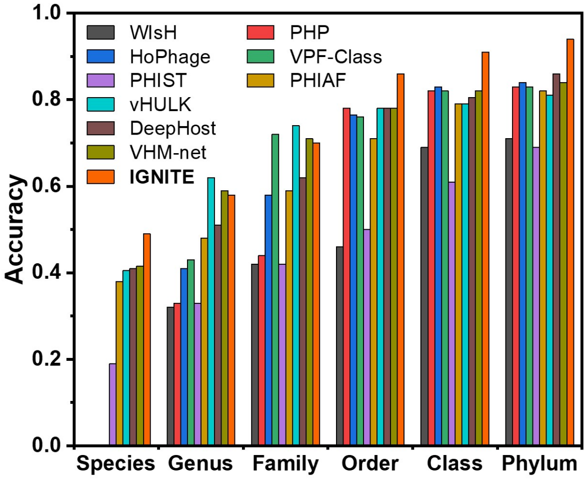

Figure 8: The comparison of the results of our method other 9 models

species to phylum on training dataset (shown in figure 7). Given the protein sequence of each virus, the model returns the prediction score of each bacteria, making the host prediction with the highest prediction score. Figure 7 shows that our model achieved promising accuracy on training dataset, especially with an accuracy of 0.89 on species level host prediction.

To further validate the robustness of our model IGNITE, we settled the parameters and applied the trained model to the test dataset for host prediction of unseen data. As shown in figure 8, our model almost outperformed 9 other models on the test virus host prediction over distinctive taxonomic levels. IGNITE achieved an accuracy of 0.49 on species level host prediction, accomplishing an improvement of 6 percent compared with the benchmark model VHM-net.

Conclusion

Recently, conventional antibiotic therapy has struggled to perform well in the treatment of bacterial infections due to the rapid development of bacterial resistance to antibiotics. In contrast, phage therapy is considered to be a promising treatment approach because of its unique ability to target and kill bacteria. Yet few phage-bacteria interactions are known, and verifying phage-bacteria interactions through extensive experiments is time and money consuming, thus is not feasible. Researchers have developed a series of host prediction methods, but as shown in figure 8, have not achieved high accuracy. In the current study, we implemented IGNITE, a deep-learning model for precise virus-host interaction prediction that utilized transformer combined with multi-layer perceptron for high-dimensional feature derivation from protein sequences of phages and and DNA sequences of hosts, followed by a decoder for binary classification. Utilizing self attention mechanism in transformer renders evaluation of importance and association of virus protein clusters, which helps distinguish the characterization of different viruses. Moreover, by concatenating the embedding of hosts and randomly selected viruses to form a negative set, the model witnesses a variety of virus-prokaryote combinations, which in turn results in a stronger ability to discern real interaction patterns. Our experimental results revealed that IGNITE achieved performance comparable to the state-of-the-art methods at genus level and beyond, outperformed 9 existing virus-host interaction prediction tools at species level host prediction and improved species level prediction accuracy on benchmark dataset by 6 percent.

References

Ahlgren, N. A.; Ren, J.; Lu, Y. Y.; Fuhrman, J. A.; and Sun, F. 2017. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research* 45(1):39–53.

Amgarten, D.; Iha, B. K. V.; Piroupo, C. M.; da Silva, A. M.; and Setubal, J. C. 2020. vhulk, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *bioRxiv*.

Breitbart, M., and Rohwer, F. 2005. Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* 13(6):278–284.

Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J. M.; Segall, A. M.; Mead, D.; Azam, F.; and Rohwer, F. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* 99(22):14250–14255.

Buchfink, B.; Xie, C.; and Huson, D. H. 2015. Fast and sensitive protein alignment using diamond. *Nature methods* 12(1):59–60.

Burstein, D.; Sun, C. L.; Brown, C. T.; Sharon, I.; Anantharaman, K.; Probst, A. J.; Thomas, B. C.; and Banfield, J. F. 2016. Major bacterial lineages are essentially devoid of crispr-cas viral defence systems. *Nature communications* 7(1):1–8.

Enright, A. J.; Van Dongen, S.; and Ouzounis, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* 30(7):1575–1584.

Galiez, C.; Siebert, M.; Enault, F.; Vincent, J.; and Söding, J. 2017. Wish: who is the host? predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33(19):3113–3114.

González-Tortuero, E.; Krishnamurthi, R.; Allison, H. E.; Goodhead, I. B.; and James, C. E. 2021. Comparative analysis of gene prediction tools for viral genome annotation. *bioRxiv*.

Gregory, A. C.; Zayed, A. A.; Conceição-Neto, N.; Temperton, B.; Bolduc, B.; Alberti, A.; Ardyna, M.; Arkhipova, K.; Carmichael, M.; Cruaud, C.; et al. 2019. Marine dna viral macro-and microdiversity from pole to pole. *Cell* 177(5):1109–1123.

Guerin, E.; Shkoporov, A.; Stockdale, S. R.; Clooney, A. G.; Ryan, F. J.; Sutton, T. D.; Draper, L. A.; Gonzalez-Tortuero, E.; Ross, R. P.; and Hill, C. 2018. Biology and taxonomy of crass-like bacteriophages, the most abundant virus in the human gut. *Cell host & microbe* 24(5):653–664.

Leite, D. M. C.; Brochet, X.; Resch, G.; Que, Y.-A.; Neves, A.; and Peña-Reyes, C. 2018a. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC bioinformatics* 19(14):151–159.

Leite, D. M. C.; Lopez, J. F.; Brochet, X.; Barreto-Sanz, M.; Que, Y.-A.; Resch, G.; and Pena-Reyes, C. 2018b. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1818–1825. IEEE.

Li, M.; Wang, Y.; Li, F.; Zhao, Y.; Liu, M.; Zhang, S.; Bin, Y.; Smith, A. I.; Webb, G. I.; Li, J.; et al. 2020. A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18(5):1801–1810.

Liu, D.; Ma, Y.; Jiang, X.; and He, T. 2019. Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. *BMC bioinformatics* 20(16):1–10.

Lu, C.; Zhang, Z.; Cai, Z.; Zhu, Z.; Qiu, Y.; Wu, A.; Jiang, T.; Zheng, H.; and Peng, Y. 2021. Prokaryotic virus host predictor: a gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC biology* 19(1):1–11.

Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E.; Goodfellow, I.; Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; et al. 2018. Advances in neural information processing systems.

Ruohan, W.; Xianglilan, Z.; Jianping, W.; and Shuai Cheng, L. 2022. Deephost: phage host prediction with convolutional neural network. *Briefings in Bioinformatics* 23(1):bbab385.

Shang, J., and Sun, Y. 2021. Predicting the hosts of prokaryotic viruses using gcn-based semi-supervised learning. *BMC biology* 19(1):1–15.

Wang, W.; Ren, J.; Tang, K.; Dart, E.; Ignacio-Espinoza, J. C.; Fuhrman, J. A.; Braun, J.; Sun, F.; and Ahlgren, N. A. 2020. A network-based integrated framework for predicting virus–prokaryote interactions. *NAR genomics and bioinformatics* 2(2):lqaa044.

Zhang, M.; Yang, L.; Ren, J.; Ahlgren, N. A.; Fuhrman, J. A.; and Sun, F. 2017. Prediction of virus-host infectious association by supervised learning methods. *BMC bioinformatics* 18(3):143–154.