Genome Analysis

Identification of bacteriophage genome sequences with representation learning

Zeheng Bai¹, Yao-zhong Zhang¹,*¹, Satoru Miyano¹,², Rui Yamaguchi¹,³,⁴, Kosuke Fujimoto⁵,⁶, Satoshi Uematsu⁵,⁶, and Seiya Imoto¹,⁶,*¹

¹Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan
²M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan
³Division of Cancer Systems Biology, Aichi Cancer Center Research Institute, Nagoya, Japan
⁴Division of Cancer Informatics, Nagoya University Graduate School of Medicine, Nagoya, Japan
⁵Division of Metagenome Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan
⁶Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

* To whom correspondence should be addressed.

Abstract

Motivation: Bacteriophages/Phages are the viruses that infect and replicate within bacteria and archaea, and rich in human body. To investigate the relationship between phages and microbial communities, the identification of phages from metagenome sequences is the first step. Currently, there are two main methods for identifying phages: database-based (alignment-based) methods and alignment-free methods. Database-based methods typically use a large number of sequences as references; alignment-free methods usually learn the features of the sequences with machine learning and deep learning models.

Results: We propose INHERIT which uses a deep representation learning model to integrate both database-based and alignment-free methods, combining the strengths of both. Pre-training is used as an alternative way of acquiring knowledge representations from existing databases, while the BERT-style deep learning framework retains the advantage of alignment-free methods. We compare INHERIT with four existing methods on a third-party benchmark dataset. Our experiments show that INHERIT achieves a better performance with the F1-score of 0.9932. In addition, we find that pre-training two species separately helps the non-alignment deep learning model make more accurate predictions.

Availability: The codes of INHERIT are now available in: https://github.com/Celestial-Bai/INHERIT.

Contact: yaozhong@ims.u-tokyo.ac.jp and imoto@hgc.jp

Supplementary information: Supplementary data are available at BioRxiv online.

1 Introduction

Bacteriophages (phages for short) are the viruses that infect bacteria and archaea, and are rich in human body (Fuhrman, 1999; Edwards and Rohwer, 2005; Rohwer and Thurber, 2009; Rodriguez-Valera et al., 2009; Reyes et al., 2012). To study the role of phages in microbial community in the human body, we need to identify phages from the metagenome nucleotide sequences (Marquet et al., 2020; Fang et al., 2019). Many methods have been proposed to identify phages, such as VIBRANT (Kieft et al., 2020), VirSorter2 (Guo et al., 2021), Seeker (Auslander et al., 2020) and DeepVirFinder (Ren et al., 2020). We categorize them into two groups: database-based (alignment-based) methods (VIBRANT and VirSorter2), and alignment-free methods (Seeker and DeepVirFinder). Both types have their advantages and disadvantages, and they are complementary. Database-based approaches are commonly based on multiple sequence alignment (Edgar and Batzoglou, 2006; Hyatt et al., 2010; Chatzou et al., 2016) with Profile Hidden Markov Models (Eddy, 1998), which can achieve good prediction performance. However, such prediction speed is generally limited by alignment. Alignment-free methods usually can predict fast. However, subject to the training process of the machine learning and deep learning models, we need to balance the amount of phage and bacteria data (Japkowicz and Stephen, 2002), which affects the amount of information obtained. The introduction of them can be found in Section 1.1, Supplementary Methods.

In proposing the MSA Transformer, Rao et al. (2021) demonstrated that pre-trained Transformer-based models can have comparable performance to HMM Profiles and are even better in some cases. That indicates the core of the database-based approaches, HMM Profiles, can be realized for...
INHERIT is an integrated model based on the DNA sequence language model. DNABERT, with two pre-trained models as references. It reaches the best performance compared with four existing state-of-the-art approaches: VIBRANT, VirSorter2, Seeker, and DeepVirFinder. INHERIT outperforms them with the highest F1-score of 0.9932 in our test.

2 Methods

INHERIT is a model with DNABERT as the backbone and uses two pre-trained models as references (see the pipeline in Figure 1A). DNABERT is an extension of BERT (Devlin et al., 2018) on nucleotide sequences. The structure of BERT model contains of 12 Transformer layers (Vaswani et al., 2017), and Transformer is a neural network composed mainly of multi-head self-attention. Multi-head self-attention is a mechanism that can be expressed with (Vaswani et al., 2017).

INHERIT is an extension of BERT (Devlin et al., 2018) on nucleotide sequences. The structure of BERT model contains of 12 Transformer layers (Vaswani et al., 2017), and Transformer is a neural network composed mainly of multi-head self-attention. Multi-head self-attention is a mechanism that can be expressed with (Vaswani et al., 2017).

INHERIT is an extension of BERT (Devlin et al., 2018) on nucleotide sequences. The structure of BERT model contains of 12 Transformer layers (Vaswani et al., 2017), and Transformer is a neural network composed mainly of multi-head self-attention. Multi-head self-attention is a mechanism that can be expressed with (Vaswani et al., 2017).

INHERIT is an extension of BERT (Devlin et al., 2018) on nucleotide sequences. The structure of BERT model contains of 12 Transformer layers (Vaswani et al., 2017), and Transformer is a neural network composed mainly of multi-head self-attention. Multi-head self-attention is a mechanism that can be expressed with (Vaswani et al., 2017).
INHERIT: identifying phages with representation learning

the "[CLS]" token is extracted and generate two outputs for each fine-tuned DNABERT through a dense layer (Wolf et al., 2020). INHERIT eventually predicts the label class based on the four outputs with a dense layer. The prediction of the whole sequence is the average of predictions of all segments, which we call the "score" of the sequence. Here we used the pre-train-fine-tune paradigm to build INHERIT (see the process of pre-training and fine-tuning INHERIT in Figure 1B). To deal with the information bias that may be caused by data imbalance (Thabtah et al., 2020), we pre-trained bacteria and phages independently. The number of bacteria we have known is much larger than the number of phages, and the length of bacteria is also longer (Chantsivili et al., 2001). If we want the pre-training set to carry a considerable amount of data, the segments belonging to the bacteria will be bound to be much more than those belonging to the phages. If we combine bacteria and phages in one pre-trained model, the model will learn much more about bacteria than phages. Therefore, we prepared two pre-trained models for INHERIT. Here, we used Masked Language Modeling (Devlin et al., 2018; Naseem et al., 2021) as the pre-training task, which is the same as Ju et al. (2021). For the detailed settings of the pre-training, please refer to Supplementary Methods Section S1.2.

Fine-tuning is very similar to the traditional training strategy (Dodge et al., 2020). The difference between them is that traditionally we randomly initialize the deep learning model before we start to train the model. However, we will transfer most of the pre-trained model weights as the initialization before we start to fine-tune the deep learning framework. Here we transferred the weights of non-linear-probing layers of the two pre-trained models to initialize INHERIT. The linear layers of INHERIT were still randomly initialized because we could not transfer the weights from the pre-trained models. All model parameters were fine-tuned together with a balanced training set. For hyperparameters and platforms of fine-tuning, please see Supplementary Methods Section S1.2.

2.2 Datasets

2.2.1 Pre-training sets

To make pre-trained models carry as much biological information as possible, we pre-trained bacteria and phages separately and did not balance the size of the two pre-training sets. For the bacteria pre-training set, we used ncbi-genome-download (https://github.com/kblin/ncbi-genome-download) to obtain the complete bacteria genome sequence from the NCBI FTP. We used the command ncbi-genome-download-formats fasta --assembly-levels complete bacteria. All of those bacteria sequences were high quality, and we called them "bacteria assemblies". We chose 718,663 segments. We also chose 2,643 sequences not in the pre-training set as the validation set, generating 186,121 segments. However, we did not have as many sequences to choose from for phages, so we selected 10,574 phage sequences from the pre-training set that possessed a quality comparable to the bacteria assemblies, generating 718,663 segments. We also chose 2,643 sequences not in the pre-training set as the validation set, generating 186,121 segments.

2.2.2 Training set and validation set for fine-tuning

For bacteria during fine-tuning, we randomly selected 260 bacteria sequences that were not in the pre-training and test sets but bacteria assemblies. 217 bacteria sequences were used as the training set, generating 718,879 segments, and the remaining 43 were used as the validation set, generating 188,149 segments. However, we did not have as many sequences to choose from for phages, so we selected 10,574 phage sequences from the pre-training set that possessed a quality comparable to the bacteria assemblies, generating 718,663 segments. We also chose 2,643 sequences not in the pre-training set as the validation set, generating 186,121 segments.

2.2.3 Test set for comparisons

The test set we use is one of the third-party benchmark tests previously proposed by Ho et al. Ho et al. (2021) for virus identification methods, called the ReSeq test set. Since our method identifies phages and not other viruses, we only used data related to phages. The ReSeq test set contains 710 bacteria sequences and 1,028 phage sequences. However, since there are 19 bacteria sequences removed from NCBI RefSeq database, we used the rest of them, including 691 bacteria sequences and 1,028 phage sequences, to examine the performance of the methods on phage identification. IT should be added that, in that article (Ho et al., 2021), the authors split the sequences in this test set into 1kb to 15kb segments on average and predict the results and calculate metrics on segment level to make a benchmark test. However, since we consider INHERIT to determine whether a sequence is a phage or not in applications, we compared INHERIT with other existing methods on the sequence level. For experimental details and results, please refer to Section 3.1.

We have posted the accessions of the sequences used in each dataset and the sources we obtain in Supplementary Table S1.

3 Experiments

3.1 Benchmarking INHERIT with VIBRANT, VirSorter2, Seeker, DeepVirFinder

In this section, we compared INHERIT with four state-of-the-art methods: VIBRANT, VirSorter2, Seeker, and DeepVirFinder, which are the representatives of database-based and alignment-free methods. We used a third-party benchmark dataset to conduct experiments and analyses.

3.1.1 Experimental setups

Baselines: Until our work is completed, INHERIT is the only model that integrates the features of both database-based and alignment-free approaches. Thus, we chose two representatives which have achieved state-of-the-art from each of the two methods, VIBRANT (Kieft et al., 2020) and VirSorter2 (Guo et al., 2021); and Seeker (Auslander et al., 2020) and DeepVirFinder (Ren et al., 2020), to compare it with INHERIT. To ensure the comparison is as fair as possible, we used each method’s default commands for predictions as much as possible, and we set corresponding rules for some methods to ensure that the prediction formats of each method are as consistent as possible. Detailed settings for each method can be found in Supplementary Methods Section S1.3.

Evaluation metrics: The evaluation metrics we chose are:

\[
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN} \\
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \\
\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
and AUROC (Area Under the Receiver Operating Characteristic curve) and AUPRC (Area Under the Precision-Recall Curve). In this paper, TP is the number of phage sequences successfully identified as phages, while FN is the number of bacterial sequences incorrectly identified as phages. TN is the number of bacterial sequences successfully identified as bacteria and FP is the number of phage sequences incorrectly identified as bacteria. AUROC and AUPRC are calculated based on the score of each sequence and the real value (phages are recorded as 1 and bacteria as 0). Here we calculate all evaluation metrics by scikit-learn v1.0.2 (Pedregosa et al., 2011).

### Prediction results

The predictions of VIBRANT, VirSorter2, Seeker, DeepVirFinder, and INHERIT for all the sequences in the test set can be seen in Supplementary Table S2.

#### 3.2 Ablation study

INHERIT performs better than existing methods: From the results (see Table 1), compared to VIBRANT, VirSorter2, Seeker, and DeepVirFinder, INHERIT performs better than other existing methods. Moreover, the overall performance of INHERIT is an order of magnitude more precise. Even if we use the default threshold of 0.5, the recall of INHERIT does not differ much from that of VIBRANT. Significantly, the high F1-score of INHERIT proves that INHERIT performs well enough when we use the default parameters and is competent for most application scenarios. From the p-values of the DeLong Test (DeLong et al., 1988) of every two methods, the ROC curves for each of the two approaches are statistically significantly different. Therefore, the high AUROC and AUPRC indicate that INHERIT can distinguish between phages and bacteria better.

The performance of INHERIT is less affected by the length of nucleotide sequences: We find the performance of INHERIT is less sensitive to the length of genome sequences compared with other methods. We divided the phage and bacteria into five length intervals, each based on the quartiles of lengths in the test set. The length intervals are according to the following rules: “minimum - first quartile”, “first quartile - median”, “median - third quartile”, “third quartile - maximum of non-outliers”, “the maximum of non-outliers - outliers”. For phage sequences, there are “less than 42,000 bp”, “42,000 bp - 51,000 bp”, “51,500 bp - 91,600 bp”, and “greater than 91,600 bp”. For bacteria sequences, there are “less than 2,788,000 bp”, “2,788,000 bp - 4,107,500 bp”, “4,107,500 bp - 4,938,000 bp”, and “greater than 4,938,000 bp”. We calculated the true positive rates of VIBRANT, VirSorter2, Seeker, DeepVirFinder, and INHERIT in the four intervals of phage and the true negative rates in the four intervals of bacteria, respectively. That allows exploring whether these five methods perform consistently and robustly at different sequence lengths for both phages and bacteria. We can find that INHERIT is the only method to maintain the true positive rate of 0.99 and the true negative rate of 0.97 in different intervals (see Figure 2). It does not always perform the best in every interval, but it is more robust than other models. For example, the true negative rate of VIBRANT in the “2,788,000 bp - 4,107,500 bp” interval is 0.994, which is better than INHERIT (0.977), but its performance in the “4,938,000 bp - 8,000,000 bp” interval is only 0.767.

INHERIT has appropriate prediction speed: INHERIT is still a model based on a deep learning framework, so it does not take as long a time to predict as database-based approaches. We calculated the average time required for VIBRANT, VirSorter2, Seeker, DeepVirFinder, and INHERIT to predict phage sequences and bacteria sequences in the test set. The results (see Table 2) show that the predictions of VIBRANT and VirSorter2 take much more time than Seeker and INHERIT. From our experiment, VIBRANT and VirSorter2 even cannot predict the whole bacteria test set at once. That indicates that even though database-based methods perform better than alignment-free methods, they consume a long time to predict and are more sensitive to dataset size. However, INHERIT has high performance and predicts the second fastest among VIBRANT, VirSorter2, Seeker, and DeepVirFinder. Although INHERIT takes a longer time to predict than Seeker, it takes much less time than DeepVirFinder. DeepVirFinder selects different models for prediction based on the length of the target sequence. In our experiments, we offered the same environment to INHERIT, Seeker, and DeepVirFinder, but DeepVirFinder is still much slower than the other two methods, even if it is based on a convolutional neural network (Lecun and Bottou, 1998; O’Shea and Nash, 2015). We conjecture that it may be because it consumes much time in the process of DNA sequence encoding and model selection. INHERIT uses one model to make predictions for each sequence regardless of the length, showing that INHERIT can give accurate predictions in an appropriate time budget.

3.2 Ablation study

We used the validation set to do an additional experiment for considering ablation. In this section, we discuss how INHERIT’s pre-training strategy and deep learning structure affect performance. There are two main differences between INHERIT and the backbone: DNABERT. First, we used a strategy to pre-train two species independently. To use pre-training to provide references and solve imbalanced bacterial and phage datasets, we pre-trained bacteria and phages separately. In addition, we utilized two pre-trained models and have them fine-tuned simultaneously. This model structure has twice the number of parameters as DNABERT. To explore whether pre-training would help deep learning frameworks improve performance, we made an ablation study on the validation set. In training INHERIT, we first trained two pre-trained models on two separate datasets of different sizes and species. Then, the weights of these two pre-trained models were
**INHERIT: identifying phages with representation learning**

Compared to INHERIT (w/o pre-train), INHERIT significantly improves models with each other, and these three curves are statistically significant. We also made DeLong Test for these three can give INHERIT a further boost in very high accuracy without increasing INHERIT, allowing it to outperform DNABERT. However, pre-training whole genome sequence. The increase in the number of parameters helps performance on the segment level, the more accurate the prediction of the level. Meanwhile, although we trained on segment level, the better of the metrics on sequence level and performs the best on a segment all metrics and performs better than INHERIT (w/o pre-train) in most 3), INHERIT improves on the excellent performance of DNABERT in segment in Supplementary Table S3. Based on the results (see Table 3).

Pre-training independently will allow the deep learning framework to give predictions closer to the true values. We plotted boxplots of the scores generated by DNABERT, INHERIT (w/o pre-train), and INHERIT for phage and bacteria samples in the validation set on sequence level (see Figure 3). For the vast majority of samples, the score generated by INHERIT is closer to the true value (1 for phages and 0 for bacteria). For example, for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT (w/o pre-train) gives a score of 0.4968. Neither model classify where none of the three predict correctly or all of the three predict correctly. This point even stands for the samples training provides great help for deep learning frameworks to give more correct predictions on both species. Compared with DNABERT, both two differences help INHERIT to make better performance.

We tested all models’ performance on the validation set. Here we evaluated all three models on both sequence and segment levels because the validation set is balanced on the segment level while severely imbalanced on sequence level. Values corresponding to best performance are bolded. Table 2), the average prediction speed of DNABERT is 53.7031 seconds for bacteria sequences and 2.7851 seconds for phage sequences in the test respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT prediction scores were 0.4209, 0.2908, and 0.4547, respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT (w/o pre-train), and INHERIT prediction scores were 0.2237, 0.1210, and 0.4547, respectively; for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT is closer to the true value (1 for phages and 0 for bacteria). For example, for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT (w/o pre-train) gives a score of 0.4968. Neither model classify where none of the three predict correctly or all of the three predict correctly. This point even stands for the samples training provides great help for deep learning frameworks to give more correct predictions on both species. Compared with DNABERT, both two differences help INHERIT to make better performance.

We tested all models’ performance on the validation set. Here we evaluated all three models on both sequence and segment levels because the validation set is balanced on the segment level while severely imbalanced on sequence level. Values corresponding to best performance are bolded. Table 2), the average prediction speed of DNABERT is 53.7031 seconds for bacteria sequences and 2.7851 seconds for phage sequences in the test respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT prediction scores were 0.4209, 0.2908, and 0.4547, respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT (w/o pre-train), and INHERIT prediction scores were 0.2237, 0.1210, and 0.4547, respectively; for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT is closer to the true value (1 for phages and 0 for bacteria). For example, for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT (w/o pre-train) gives a score of 0.4968. Neither model classify where none of the three predict correctly or all of the three predict correctly. This point even stands for the samples training provides great help for deep learning frameworks to give more correct predictions on both species. Compared with DNABERT, both two differences help INHERIT to make better performance.

We tested all models’ performance on the validation set. Here we evaluated all three models on both sequence and segment levels because the validation set is balanced on the segment level while severely imbalanced on sequence level. Values corresponding to best performance are bolded. Table 2), the average prediction speed of DNABERT is 53.7031 seconds for bacteria sequences and 2.7851 seconds for phage sequences in the test respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT prediction scores were 0.4209, 0.2908, and 0.4547, respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT (w/o pre-train), and INHERIT prediction scores were 0.2237, 0.1210, and 0.4547, respectively; for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT is closer to the true value (1 for phages and 0 for bacteria). For example, for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT (w/o pre-train) gives a score of 0.4968. Neither model classify where none of the three predict correctly or all of the three predict correctly. This point even stands for the samples training provides great help for deep learning frameworks to give more correct predictions on both species. Compared with DNABERT, both two differences help INHERIT to make better performance.

We tested all models’ performance on the validation set. Here we evaluated all three models on both sequence and segment levels because the validation set is balanced on the segment level while severely imbalanced on sequence level. Values corresponding to best performance are bolded. Table 2), the average prediction speed of DNABERT is 53.7031 seconds for bacteria sequences and 2.7851 seconds for phage sequences in the test respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT prediction scores were 0.4209, 0.2908, and 0.4547, respectively; for bacterium NZ_CP018197, the DNABERT, INHERIT (w/o pre-train), and INHERIT prediction scores were 0.2237, 0.1210, and 0.4547, respectively; for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT is closer to the true value (1 for phages and 0 for bacteria). For example, for phage MH576062, DNABERT gives a score of 0.4707, and INHERIT (w/o pre-train) gives a score of 0.4968. Neither model classify where none of the three predict correctly or all of the three predict correctly. This point even stands for the samples training provides great help for deep learning frameworks to give more correct predictions on both species. Compared with DNABERT, both two differences help INHERIT to make better performance.

**Fig. 2:** The barplot of the true positive rates and true negative rates at different intervals. (A) shows the true positive rates for each method at different length intervals. (B) shows the true negative rate of each method at different length intervals.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>AUROC</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNABERT</td>
<td>0.9992</td>
<td>0.9224</td>
<td>0.9229</td>
<td>0.9593</td>
<td>0.9751</td>
<td>0.9996</td>
</tr>
<tr>
<td>INHERIT (w/o pre-train)</td>
<td>0.9996</td>
<td>0.9255</td>
<td>0.9263</td>
<td>0.9611</td>
<td>0.9839</td>
<td>0.9997</td>
</tr>
<tr>
<td>INHERIT</td>
<td>0.9992</td>
<td>0.9894</td>
<td>0.9888</td>
<td>0.9943</td>
<td>0.9971</td>
<td>1.0000</td>
</tr>
<tr>
<td><strong>Segment Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNABERT</td>
<td>0.8767</td>
<td>0.8809</td>
<td>0.8792</td>
<td>0.8788</td>
<td>0.9508</td>
<td>0.9507</td>
</tr>
<tr>
<td>INHERIT (w/o pre-train)</td>
<td>0.9044</td>
<td>0.8951</td>
<td>0.9008</td>
<td>0.8997</td>
<td>0.9530</td>
<td>0.9422</td>
</tr>
<tr>
<td>INHERIT</td>
<td>0.9436</td>
<td>0.9650</td>
<td>0.9539</td>
<td>0.9542</td>
<td>0.9875</td>
<td>0.9882</td>
</tr>
</tbody>
</table>

Table 3. The comparison among DNABERT (without pre-training, i.e., training from scratch), INHERIT (without pre-training), and INHERIT. We use the validation set to evaluate the performance of the three models. Here we evaluate all three models on sequence level and segment because the validation set is balanced on the segment level while severely imbalanced on sequence level. Values corresponding to best performance are bolded.
set. It is slightly faster than INHERIT. That is because INHERIT has a more complex structure than DNABERT. However, INHERIT can still give predictions faster than other methods.

4 Conclusions

In this work, we proposed INHERIT, an integrated method that combines both database-based and alignment-free approaches under a unified deep representation learning framework. It uses two pre-trained models as references and keeps the features of alignment-free methods by the deep learning structure. On a third-party benchmark dataset, we compared the proposed method with VIBRANT, VirSorter2, Seeker, and DeepVirFinder, representing database-based methods and alignment-free methods. We demonstrate that INHERIT can achieve better performance than the four existing methods in all metrics. In particular, INHERIT improves the F1-score from 0.9787 to 0.9932. Meanwhile, we also prove that using an independent pre-training strategy can make deep learning models make better predictions on both species.

Acknowledgements

The super-computing resource was provided by the Human Genome Center, Institute of Medical Science, The University of Tokyo (https://gc.hgc.jp/en/). Y.Z.’s appreciation also goes to Google who provided GCP based on the partnership between UTokyo and Google.

Conflict of Interest: none declared.

The data underlying this article are available in NCBI Nucleotide database at https://www.ncbi.nlm.nih.gov/nuccore/, and the accessions of the sequences used in each dataset can be found in online supplementary materials.

References

INHERIT: identifying phages with representation learning


