Pathogenic virus detection method based on multi-model fusion

Dr. Xiaoyong Zhao
Department of Information Management institute
Beijing Information Science and Technology University
Beijing, China
zhaoxiaoyong@bistu.edu.cn

Ms. Jingwei Wang
Department of Information Management institute
Beijing Information Science and Technology University
Beijing, China
wang-jing-wei@hotmail.com

***Abstract*—Identifying viruses is crucial for pandemics. Detecting pathogenic viruses is difficult and usually needs to spend lots of time. Here, we propose a multi-model fusion method for viruses pathogenic detection(MMFPV). We use CatBoost, Reverse-complement Convolutional Neural Networks, and, Reverse-complement Long Short-Term Memory(RC-LSTM) as base models, then, using stacking to combine these models. This method directly detects pathogenic viruses from next-generation sequencing data without virus databases. The result shows that this novel detection method has good capabilities.**

Keywords—next-generation sequencing; multi-model fusion; pathogenic viruses; deep learning

I. INTRODUCTION

As humans continue to explore the natural world, more and more unknown viruses appear, and with global warming, the dormant viruses gradually recover. For example, the outbreaks of the Ebola virus, Zika virus, and 2019 novel coronavirus(COVID-19), in recent years, have shown that the risk of outbreaks needs to be highly valued.

Currently, there are four commonly used detection methods for viruses and other microorganisms, which are Microbiological culture, polymerase chain reaction (PCR), mass spectrometry, and, next-generation sequencing (NGS), among which, NGS has theoretical advantages in the discovery of new pathogens, but its analysis algorithms are still insufficient. The current methods highly depend on the similarity of data in the existing virus pathogen database, such as BLAST and other alignment-based algorithms. Once data has low similarity or completely dissimilar, the commonly used detection methods cannot show detective results. In addition, the cross-species transmission of viruses also brings difficulties in virus identification.

Li et al. proposed a comparative study of predicting viral hosts based on alignment or alignment-free methods[1]. The results show that when alignment methods cannot detect or need to spend lots of time, the support vector machines model is a substitute method, which produces good prediction results. Zhang et al. proposed a method to quickly identify human viruses[2]. In this study, they used complete viral genome data to train several different machine learning models and investigated the effect of different lengths of k-mer on the performance of several models. Applying contigs of different lengths into the algorithm, the results show that the learner trained with complete viral genome data also has a good prediction effect on metagenomic data. In recent years, thanks to end-to-end feature extraction capabilities, deep learning has been applied to virus detection. Mock et al. proposed a study based on deep learning to predict virus hosts[3]. In the study, long short-term memory neural networks and convolutional neural networks achieve good results. Ren et al. proposed a deep learning method to identify viruses from metagenomic data. they considered the double-stranded structure of DNA and took each branch of DNA as input[4]. Tampuu et al. proposed a study to identify viruses from metagenomic data[5]. In their study, it contains Siamese Networks, and each brand of the Siamese Networks has a different meaning that one represents frequency and the other represents patterns.

In response to the problems and deficiencies of current research, we propose a Multi-Model Fusion method to detect pathogenic viruses(MMFPV) which combines k-mers, protein, and, end-to-end features into a multi-model pipeline, thus, predicting whether a gene sequencing is pathogenic to humans.

II. DATA

A. Source

On March 15, 2020, we first accessed the Virus-Host Database(VHDB) and downloaded all available data, containing 14499 items of viruses metadata; then through these metadata, we downloaded genomic data from Nation Center for Biotechnology Information(NCBI). One needs to note that one item of meta-data may have a one-to-many relationship in NCBI. In the subsequent analysis, each sequence is regarded as one item of data. The selected viruses include both RNA viruses and DNA viruses, and the sequence of the RNA viruses is encoded in the form of DNA in the reference sequence.

B. Define labels

A virus may infect both humans and other creatures. In this case, as long as the host information provided by VHDB contains "homo", the data are classified as viruses that can infect humans. 1623 viruses that can infect humans, and these viruses are set as positive labels. Compared with positive labels, the design of negative labels is relatively complicated. Considering the perspective of broadness and pertinence, we constructed two negative label datasets.

The first negative label dataset contains all viruses that cannot infect humans. Such a dataset cover a wider range of genomic data, but it may misclassify potentially dangerous mammalian or avian viruses; in addition, it also contains the bacteriophages, which may also reduce the detection performance.

The second negative label dataset only considers non-human viruses that can infect the chordate as negative labels. Studies have shown that the evolutionary path of viruses is: other animal viruses -> zoonoses -> human obligate viruses, so it is more targeted to distinguish between viruses that can only infect the chordate and humans[6].

The two datasets finally constructed are shown in Table 1 below.

III. METHOD

A. Framework

The framework of overall algorithm is shown in the Fig. 1. In the training phase, first, we extract k-mers and physicochemical properties of proteins from the genomic data, and train gradient boosting trees; then, we generate and recode date as networks input; afterwards the three base classifiers are fused to obtain the final model. In the prediction phase, viral genome sequencing data input to the model; then the model will give the prediction result for the data.

B. Feature Extraction

To train models, some features must be extracted from the genomic sequence. k-mer based feature extraction method is a common approach when using machine learning analyses genomic data. Generate from four base pairs (A, T, C, G) genomic sequence fragments of length k; then count the frequency of genomic sequence fragments of length k in the genomic sequence. When using the k-mer-based feature extraction method, as the k value increases, the number of extracted features will grow exponentially i.g. when k = 8, 65,536 features can be extracted, but, too many features as input will cause the curse of dimensionality and the performance of the learning algorithm will be significantly reduced, which will affect the prediction results; therefore, in this paper, we only extract the features when k=1,2,3,4, a total of 340 features.

TABLE I. The introduction of two datasets.

Datasets name	Positive label	Negative label	Quantity
All-non-human	Viruses that can infect humans (11,265)	viruses that cannot infect humans (273,331)	284,596
Chordata	Viruses that can infect humans (11,265)	viruses that can infect the chordate (32,392)	43,657

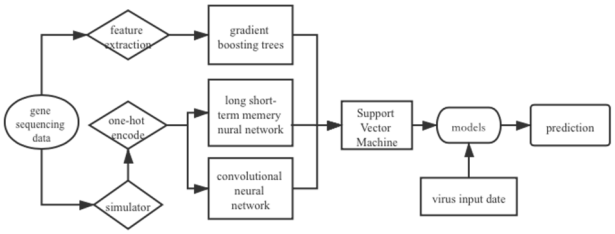


Fig. 1. The framework of overall algorithm

Proteins are a major component of viruses, and they are relatively conservative during evolution. When viruses infect cells, they use the cell's metabolic system to successively transcribe and translate the proteins needed by the virus. Therefore, the physical and chemical properties of proteins are selected as features. First, we download the protein physicochemical data from the AAindex database, and then through correlation analysis, select the 90 physicochemical properties with the lowest correlation from 554 properties. By calculating the frequency of single peptide and the information provided by AAindex, 90 physical and chemical properties of proteins are scored.

Deep learning models do not require manual design to extract features, but they still need to convert the original genomic data into a form that the models can process. Although some scholars have successfully mapped DNA sequences into genomic data in the form of pictures and used it as input for deep learning research[7], for implementing reverse-complement networks, We use a method based on one-hot encoding of every nucleotide in the sequence, reversing the sequence tensor along both axes results in the reverse-complement. In order to apply genomic data into networks, we use mason simulator to generate paired-end reads data and the length is 250dp[8]. Note that though real sequencing reads may vary in cases to cases, 250dp is a common range on Illumina platform.

The gradient boosting tree algorithms(GBDT) are widely used in academia, industry, and competitions for its state-of-the-art performance in many machine learning tasks and are almost the gold standard for structured datasets. Researchers conducted a comparative study on the most advanced GBDT packages , and the results showed that for datasets with a large number of features, especially categorized features, Catboost performs best in speed and accuracy [9].

Shrikumar et al. proposed an idea of reverse-complement network for genomic data, which can simultaneously read double-stranded DNA strands [10]. Based on the network structure, Jakub et al. studied the detection method of the novel pathogenic DNA viruses [11], and proposed an interpretable learning method for detecting novel human viruses from genome sequencing data [12].

On account of the researches above and the purpose of this article: based on the viral genome sequence to predict whether viruses is pathogenic to human, we determine two types of models as base classifiers, which are deep learning based models and CatBoost model. Considering that the selected features of the two types of models are different, fusing these two types of base classifiers complement each other to the final prediction. In addition, from the perspective of the interpretable model CatBoost model is easier than deep learning. Therefore, it was finally decided to use the CatBoost and the reverse complementary neural network structures RC-CNN and RC-LSTM for the deep learning based classifier.

IV. EXPERIMENT

The experimental environment of this article is an Alib a- ba Cloud GPU server, 4-core CPU memory 15G, NVIDI A TESLA T4 GPU, and the operating system is Ubuntu 1 6.04. Use python3.7, bio-python, keras2 backend tensorflo w1.15 and other packages.

A. *Parameters*

The data sets used in this article are shown in the data section. The parameters of each base model initially use the default parameters recommended by the algorithm author or related papers. All subsequent parameters tuning are adjusted by Bayesian optimizer. Before the deep learning, we use mason simulator to generate balanced proportion datasets. Both RC-CNN and RC-LSTM use batch training. The loss and accuracy of each batch are shown in Fig. 2. and Fig. 3. Finally, the best performing batch model is selected.

B. *Results*

For having a well-performing second layer model in stacking, we compare three simpler algorithms based on former experience, which are Logistic Regression, Support Vector Machine, and, Naive Bayesian. The comparing results on All-non-human dataset is shown in Table II, on Chordata dataset is shown in Table III.

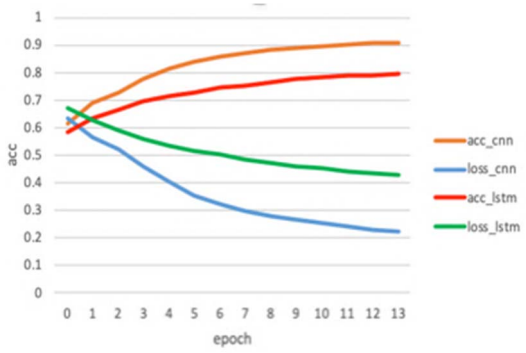


Fig. 2. The performance of RC-CNN and RC-LSTM batch training on All-non-human datasets

This research was funded by the Beijing Municipal Education Commission of Science and Technology Plan Project (KM202011232004, SM202011232008, Z2019022).

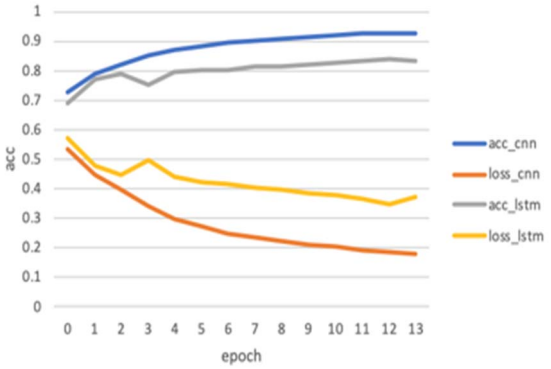


Fig. 3. The performance of RC-CNN and RC-LST M batch training on Chordata datasets

TABLE II. three algorithms comparing results on All-non-human dataset.

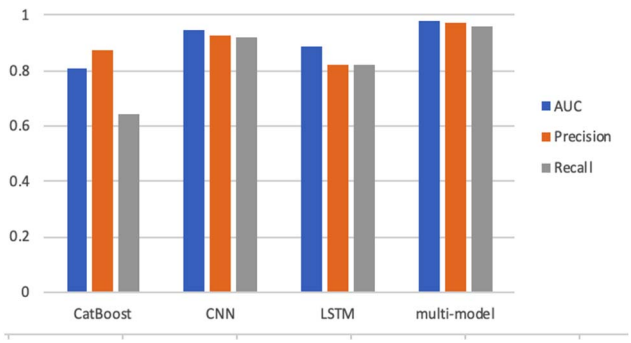
models	Precision	Recall	F1-score
Logistic Regression	0.96	0.97	0.96
Support Vector Machine	0.96	0.97	0.97
Naive Bayesian	0.95	0.96	0.95

TABLE III. three algorithms comparing results on Chordate dataset.

models	Precision	Recall	F1-score
Logistic Regression	0.92	0.86	0.89
Support Vector Machine	0.92	0.91	0.91
Naive Bayesian	0.88	0.92	0.90

When training the CatBoost classifier, 70% of the dataset is selected as the training set, and 30% as the verification set. When training deep learning classifiers, 90% of them are selected as the training set, and 10% as the verification set. Finally, the dataset split method used in CatBoost training is the same as the second layer model. Contrasting base classifiers with the fusion model showing in Fig. 4. Fig. 5. , after model fusion, the model's AUC, Precision and Recall value has been improved to a certain extent.

Fig. 4. The performance of base classifiers and fusion model on All-non-human dataset.



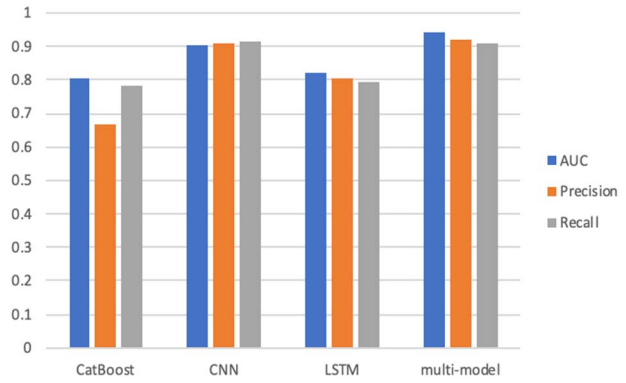


Fig. 5. The performance of base classifiers and fusion model on Chordata dataset.

C. Conclusion

In recent years, epidemics have continued to break out. Global pandemic COVID-19, which brings tremendous economics and health damage to the whole world. It gives people an illuminate that the ability to detect new viruses and recognize pathogenic viruses in time has become more and more important. The rapid development of NGS provides a massive source of big data for microbiology research, and artificial intelligence-based technology can mine potential laws from the massive big data, and discover the theoretical basis of diagnosis. Commonly used methods to detect pathogenic viruses are accurate, but it highly relies on existing viruses databases and needs to take a long time. In response to these problems, this paper proposes a pathogenic virus detection method based on multi-model fusion(MMFPV). This method integrates three kinds of features, which are end-to-end features, k-mers features, protein features, into a multi-model pipeline. Experimental results show that this method can effectively predict the pathogenic viruses. However, multi-model fusion requires a certain degree of computing resources. Without affecting the effect of the model, how to use knowledge distillation and other methods to compress the model to reduce computing resources is the future research direction. Also, further improving the interpretability of the model is also the content that needs further research.

REFERENCES

[1] Li, H., & Sun, F. (2018). Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. *Scientific reports*, *8*(1), 1-9.

[2] Zhang, Z., Cai, Z., Tan, Z., Lu, C., Jiang, T., Zhang, G., & Peng, Y. (2019). Rapid identification of human-infecting viruses. *Transbound-ary and emerging diseases*、第一届全国生物信息与传染病交叉论坛摘要集 *66*(6), 2517-2522.

[3] Mock, F., Viehweger, A., Barth, E., & Marz, M. (2019). Viral host prediction with deep learning. *bioRxiv*, 575571.

[4] Mock, F., Viehweger, A., Barth, E., & Marz, M. (2019). Viral host prediction with deep learning. *bioRxiv*, 575571.

[5] Tampuu, A., Bzhalava, Z., Dillner, J., & Vicente, R. (2019). ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PloS one*, *14*(9), e0222271.

[6] Sompayrac Lauren Sompayrac, How pathogenic Viruses work, September 2002.

[7] Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, *20*(7), 389-403.

[8] Holtgrewe, M. (2010). Mason–A Read Simulator for Second Generation Sequencing Data. *Technical Report FU Berlin*.

[9] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

[10] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Reverse-complement parameter sharing improves deep learning models for genomics. *bioRxiv*, 103663.

[11] Bartoszewicz, J. M., Seidel, A., Rentzsch, R., & Renard, B. Y. (2020). DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, *36*(1), 81-89.

[12] Bartoszewicz, J. M., Seidel, A., & Renard, B. Y. (2020). Interpretable detection of novel human viruses from genome sequencing data. *BioRxiv*.