International Core Journal of Engineering Volume 8 Issue 7, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0039

Advances in Machine Learning-based Prediction of Viral Hosts

Jiangbo Tong, Lingzhi Hu*

Shaanxi University of Traditional Chinese Medicine, China

Abstract

Viruses have caused incalculable damage to humans due to their rapid mutation ability. With the development of computer-based technology, more and more computational methods for predicting hosts have been developed to solve the host identification prediction problem and provide theoretical basis. These methods can be used to rapidly predict the hosts of virus interactions and to make decisions about epidemic control and prevention.

Keywords

Viruses; Machine Learning; Host Identification.

1. Preface

Viruses are non-cellular organisms that need to attach to specific plants, animals or microorganisms in order to survive, but sometimes under certain conditions, crossing the host barrier and infecting other organisms can occur, subsequently causing serious public health problems such as influenza viruses, which belong to the orthomyxoviridae family in the taxonomy of viruses. The genome is segmented, single-stranded, negative-stranded RNA[1] The genome is segmented, single-stranded, negative-stranded RNA. Because the genome is segmented, it is prone to gene reassortment between different strains of the same type. In particular, the HA gene of human influenza A viruses is subject to frequent point mutations, resulting in amino acid sequence substitutions in the HA protein molecule that encodes it, causing frequent antigenic drift, so that each antigenic drift often results in a different level of influenza epidemic and some influenza viruses exhibit high virulence and require more rigorous experimental treatment in biological laboratories. The study of the mechanisms of virus-host interaction is particularly important for influenza preparedness.

Humans know too little about known virus-host relationships. The gap between the number of prokaryotic viruses currently available by sequencing and the number of known virus-host relationships is rapidly widening. Traditional experimental approaches are not only expensive and time consuming, but worse still, direct virus-host relationships are rarely detected as less than 1% of microbial hosts can be successfully cultured in the laboratory. Therefore, there is an urgent need for a computational method that directly pairs predicted hosts.

Machine learning has been around since the 1970's. With the development of computer technology and the improvement of computer hardware, machine learning can make many complex tasks and data processing, and with the rise of deep learning in recent years, more complex and non-linear tasks can be made. Deep learning is built on the basis of artificial neural networks. Where an artificial neural network is an algorithmic mathematical model for distributed parallel processing of data modelled on the way the brain is neurally connected in higher organisms. This network model is a model capable of modelling complex logic and establishing non-linear relationships by adjusting the interconnections between a large number of internal nodes to change the complexity of the system for the purpose of processing complex information[2] The rapid development of this technology has led to the application of neural networks playing an important role in a wide range of industries.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0039

2. Traditional Experimental Methods for Identifying Virus-host Interactions.

2.1 Prediction of Host Identification based on Yeast Two-hybrid Crosses

Yeast two-hybrid is used to predict whether a virus interacts with its host by is using a particular viral protein as a bait to interact with a known host protein. Although yeast two-hybrid methods have the advantages of ease of use, versatility and high throughput, there are problems with false negatives and false positives when using yeast two-hybrid methods[3] . This is because certain prey-DNA structural binding domains fused to decoy-DNA transcriptional activation domains are able to self-initiate the expression of the transcriptional system without contact with each other, leading to false positive results. In addition, there may be proteins that form complexes with other proteins and subsequently activate reporter genes, resulting in false positives in yeast two-hybrid crosses. Yeast two-hybrid crosses may express proteins that are toxic to yeast and subsequently inhibit the growth of yeast or the expression of other reporter genes. In addition, the addition of substances such as aminotriazoles to the medium to suppress background expression can be toxic to yeast[4]. There are also proteins that are masked by the experimental background because of their weak interactions, and yeast-expressed proteins with fusion genes that may alter the spatial location of proteins and prevent protein interactions, resulting in false negatives.

2.2 Immunoprecipitation

Immunoprecipitation is a classical method for studying protein interaction s in vivo based on specific interactions between antigens and antibodies, detec ting protein interactions under near-natural physiological conditions, and dete cting weak or transient protein interactions[5] It can detect weak or transient protein interactions. Antibody selection is flexible, with the use of specific or tagged antibodies. Disadvantages of the assay include its unsuitability for large-scale protein interaction screening, the inability to demonstrate that t he interaction of two target proteins occurs directly, and the possible involve ment of other macromolecules in the binding process. Artificial manipulation du ring the preparation of cell lysates can also lead to physiologically unrelated protein interactions causing false negatives.

3. Machine Learning-based Host Identification Prediction.

3.1 Deep Learning Methods based on Convolutional Neural Networks.

To predict the interaction between human and viral proteins, Zhang Ziding's team at China Agricultural University combined evolutionary sequence features with a concatenated convolutional neural network architecture and a multilayer perceptron[6].This architecture outperforms various feature-coding-based machine learning and state-of-the-art prediction methods. The researchers introduced two transfer learning methods, namely freeze-type and fine-tuned-type, to accurately predict interactions in the target human-virus domain by retraining the convolutional layers in a human-virus domain-based training. Finally, the researchers used the freeze-type transfer learning method to predict the human- SARS-CoV-2 PPI, and the results showed that the prediction was topologically and functionally similar to experimentally known interactions. Transfer learning via multiscale convolutional neural layers for human-viral-protein interaction prediction demonstrates machine learning-based protein interactions that are shown to have high sensitivity and accuracy. Furthermore, migration learning can effectively apply prior knowledge gained from large source datasets or tasks to small target datasets or tasks, thereby improving prediction performance.

3.2 Machine Learning-based Screening of Ebola Hosts

For example, Simon Babayan, Richard Orton and Daniel Streicker[7] studied the genomes of over 500 viruses to train machine learning algorithms to match patterns embedded in the viral genomes to their animal origins. The models were able to accurately predict which animal host each virus came from and whether the virus needed to be transmitted by biting the blood to determine whether the vector was a tick, mosquito, gnat or whitefly. Next, the researchers applied the models to viruses whose hosts and vectors were not yet known, such as Crimean Congo haemorrhagic fever, Zika and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0039

MERS. the hosts predicted by the models often confirmed the current best guess in each field. The study found that two of the four Ebola viruses thought to have bat reservoirs actually have the same or stronger support as primate viruses, which may point to a non-human primate rather than a bat.

3.3 Prediction of Host Models based on Pairwise Convolutional Neural Networks

Yanni Sun's team at the Chinese University of Hong Kong[8] , proposed a semi-supervised learning model for host prediction of new prokaryotic viruses. A knowledge graph was constructed using virus protein similarity and virus-host DNA sequence similarity. A graph convolutional network was then trained on viruses with both known and unknown hosts to improve the sensory domain and learning ability of the model, and its performance was compared with other state- of-the-art methods designed specifically for virus host classification (VHM-net, WIsH, PHP, HoPhage, RaFAH, vHULK and VPF-Class). It is shown that the results of the model outperform other known methods, demonstrating the effectiveness of semi-supervised learning methods using graph convolutional neural networks based on them. Also, another particular advantage of the model is its ability to predict hosts from new taxa.

3.4 Predicting the Antigenicity of Viruses based on Machine Learning

Machine learning can also predict the antigenicity of viruses and characterize the antigenicity of highly virulent viruses without requiring high levels of biosecurity[9-11]. Secondly, vaccines take a long time to develop and produce and may fail due to mutations in the virus. Rapid and accurate prediction of influenza virus antigenicity is important for vaccine monitoring, screening and production, therefore, predicting influenza virus antigenicity based on computational methods not only reduces the time to detect influenza virus antigenicity, but also expands the scope of influenza surveillance and improves the efficiency of influenza vaccine screening.

4. Outlook.

Viruses are constantly mutating during the evolutionary process, gaining the ability to infect new hosts through constant mutation, and detecting the hosts of viruses in traditional experimental methods is time-consuming, costly and has the disadvantage of large errors in results. By using machine learning to predict the hosts of viruses, it is possible to quickly predict and assess the hosts of viruses for rapid outbreak prevention and control. The use of machine learning to predict the host of a virus allows the evolutionary features of its genome to be identified more quickly so that the rapid spread of the virus can be stopped. As the level of scientific research increases and research capabilities continue to improve, in the near future, more access to relevant data on virus- host protein interactions may be obtained, further allowing for the refinement of machine learning-based models that can accelerate the study of viral infection mechanisms. And it lays the foundation for drug development and the principles of virus-host interactions. In addition to this, data from conventional experimental methods can be supplemented. These suggest that machine learning- based virus-host interaction interactions will play an increasingly important role in the future.

Acknowledgments

Innovative training Project for College students of Shaanxi University of traditional Chinese Medicine in 2020: S202010726028.

References

[1] Li Xianxi, Wang Chenggui, and Zhuang Li. "Research progress on rapid diagnosis of influenza virus." Journal of Beihua University: Natural Science Edition 4.2 (2003): 6.

[2] Xiong, Kaili. Research on machine writing based on deep learning. Diss.

[3] Yu, H. Q., and Yang, W. S.. "Construction and feasibility test of yeast two-hybrid prey plasmids for VCP truncated genes." China Health Nutrition 25.013 (2015): 31-32. liu G.T., Zhang Shuyu, and Wei K.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0039

Research progress of yeast two-hybrid technology[J]. Shandong Animal Husbandry and Veterinary Medicine,2021,42(06):57-59.

[4] Shi, Jinfeng et al. "Technical advances in virus-host protein interaction studies." Advances in Animal Medicine 42.12(2021):5.

[5] Hu, X. , et al. "DeepTrio: a ternary prediction system for protein-protein interaction using mask multiple parallel convolutional neural networks." Bioinformatics (2021).

[6] Cui F, Cole HA, Clark DJ, Zhurkin VB. Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing[J]. Nucleic Acids Res, 2012 ,40(21): 10753- 10764.

[7] Babayan, Simon A , R. J. Orton , and D. G. Streicker . "Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes." science 362.6414 (2018): 577- 580.

[8] Shang, J. , and Y. Sun . "Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning."(2021).

[9] Wang Jia. Machine learning-based prediction of cross-species transmission and antigenic relationship of influenza A virus. Diss. Huazhong University of Science and Technology.

[10] Wang Jia, and Ding Xiongfei. "A study on the application of data mining in predicting the host preference of influenza A virus protein." Digital Technology and Applications 6 (2018):2.

[11] Chen WJ. Bioinformatics-based Prediction of Antigenic Variation and Antigenic Evolution of Influenza B Virus. Diss. Hunan University.