A Study in the use of Artificial Intelligence in the Identification of Unknown Viruses

Shreyas Dhake, Smital Dhake

Abstract— New viruses are always being found. This suggests that there are many more viruses in the environment that we don't know. Some of which may be harmful to humans. This raises the need of being able to detect and/or predict new, unknown viruses. Artificial intelligence consists of numerous algorithms that could overcome the limitations of traditional detection methods and assist in viral genome classification and therefore detection and prediction of unknown viruses.

Index Terms— metagenomics, virus, viral genomes, neural networks, deep learning, random forests, convolutional neural networks

1 Introduction

VIRUSES form a large percentage of the biological entities in our environment. They exist in almost every ecosystem on Earth interacting with all living things. Although viruses can be beneficial having necessary, positive effects, they can also cause infections, which can lead to life-threatening diseases. Viruses can infect animals, plants, insects as well as all other wildlife [1].

Viruses are very small parasites displaying a wide variety of shapes and sizes. Viruses can vary in length and structure as well as in the number of DNA or RNA molecules it contains which can be single-stranded or double-stranded. Viruses contain genetic material in the form of DNA or RNA. Proteins forming a coat called a capsid protect this core material. The capsid prevents the virus from being destroyed by the host's enzymes. Some viruses may have an additional protective layer called the envelope, which has spikes to allow it to attach to host cells. Usually this layer is formed by altering and modifying the cell membrane of the host. As soon as a virus is capable of causing an infection, it is called a virion [2].

A virus does not contain the cellular systems to be able to reproduce itself; they are obligate parasites. Therefore, it requires a host cell to enable it to replicate. The key role of the virus/virion is to deliver the genome into the host cells so that the genome can be transcribed and translated by the host cells. There are many different mediums of entry into the host for the virion, for example, via nasal cavity, open wound, or by binding to cell surface receptors. This release of genomes interferes with the host's cellular systems by instructing the host cells to produce viral proteins. This creates the environment in which the virus is able to spread.

- *Shreyas Dhake is currently pursuing A Levels at Watford Grammar School for Boys, UK. E-mail: shreyas.dhake@gmail.com*
- *Smital Dhake has completed master's degree in Robotics at King's College London, UK. E-mail: smitald98@gmail.com*

Since the viral nucleic acid alone is infectious, it became the most crucial characteristic in identifying and classifying viruses. The viral genome holds the complete information for the virus. This resulted in accelerated development in the application of metagenomic analysis.

2 Metagenomics

Metagenomics are a set of bioinformatic and genetic tools enabling the analysis of genomes which have been extracted from environmental samples [3]. This technique provided the identification and characterization of genome by genomic sequencing. This analysis of genetic reservoirs also allowed them to characterize novel proteins and compounds with potential biological impacts. This focuses on the genes in the sample, and how the genes function and impact other's functions. This allows scientists to define environments, and its microorganisms, biochemical and geochemical characteristics and their potential influences which cannot be achieved by traditional cultivation approaches. This means that it is possible to monitor and detect early any dangerous viral contaminants.

In the past few years, metagenome sequence datasets have grown extensively. This has reduced costs for further sequencing projects and for developing tools on sequence-based metagenomics. A general sequence-based metagenome project begins with sampling, then, DNA extraction, DNA sequencing, annotation, statistical analysis and data storage as metadata [3].

The modern sequencing techniques are named as next-generation sequencing (NGS) which have increased speed, accuracy and high throughput as compared to first generation sequencing. Additionally, reduced costs have been achieved with NGS. NGS provides deep, in-parallel DNA sequencing. This was achieved by attenuating the sequencing reactions and development in detection systems [4]. This provides a broader understanding of the structural and functional characteristics of the genomes and gives an insight into potential deadly viruses which are unknown.

3 Study & Analysis

Neural networks are a powerful algorithm used vastly as a machine learning technique in numerous applications. However, initially, metagenomics literature lacked its presence. In 2015, Ditzler et al [5] discussed and tested the standard multi-layer perceptron with two other deep learning techniques: deep belief network and recursive neural network on metagenomic data. The motivation to explore the feasibility of deep learning algorithms in metagenomics is because of the advantages it provides. Compared to traditional methods, deep learning algorithms allows features to be obtained from raw data and be able to make predictions regarding new, unknown data, simultaneously.

The begin with, the deep learning approaches were described. Supervised classification provides the benefit of being able to classify unlabeled data segments (e.g. genomes) to classes (e.g. phenotypes) as required. Additionally, the quality of the data can be analyzed by the class separability obtained after the classification of the data. Although high class separability is a desirable characteristic, the relationships between the classes need to be understood. This gives information about the structure of the microbial community present in the data sample. A graphical representation of this structure, in the form of a hierarchical tree, allows the comparison of different data samples. However, the tree structure is not learned from the input data samples. Whereas, a layered tree-like structure model will learn feature representations better from the data samples enabling unknown data predictions. This is because deep neural networks are able to converge to a more superior local optima by implementing prior training. This works better than simply backpropagation.

A deep Multi-Layer Perceptron Neural Networks (MLPNN) is a highly popular Artificial Neural Network (ANN) but training a deep MLPNN is computationally extensive and therefore expensive. In addition, the local minimum is reached very quickly because the error gradient attenuated iteratively. This gives rise to the need of deep belief networks (DBN) which are easy to understand and implement. Although there are other deep learning methods that can provide significant insights.

By using classification accuracy as a performance metric, they found the MLPNN and Random Forest Classifier (RFC) performed better than DBN. However, due to only a few parameters available in some classifiers, the choice of features becomes critical. So, if feature selection is done more effectively, it may be possible to reduce the classification losses. Similarly, in terms of the performance across various experiments, RFCs performed better. Moreover, Recursive Neural Networks (RNN) provided further advantages of representing the metagenomic samples in a hierarchical structure which gave more insight into the data. Overall, they found that deep learning techniques may not be suitable for metagenomic datasets. However, improved results may be obtained if larger datasets were used and other forms of performance metric is used, since classification accuracy does not give a complete representative.

Specifically, to viral identification, Abdelkareem et al [6] in 2018 created a deep attention model called VirNet, which overcomes the limitations of previous techniques by handling the diversity of viral genomes differently. Previous techniques had constraints, as there is no standard marker gene for all viruses. VirNet performs characterization of viral genomes from a mixture of viral and bacterial genomes. This also acts as a purification of metagenomic data from bacterial sequences. Previous tools perform similarity searches on the databases of sequences. However, these databases are nowhere close to representing the viral diversity in the environment and hence has limitations. Also, they are only able to handle small numbers of contigs which is another drawback. Gradually, the tools, such as VirFinder, became faster, efficient and more accurate however there were always some limitations such as slow processing of metagenomic data. Comparing the accuracy of VirNet and VirFinder, VirNet were found to be more accurate. This is because it is trained with viral databases that have good statistical models. This creates generalization for classification of all genomes, along with its ability to learn the suitable feature extraction of input data samples. However, the accuracy of VirNet can be improved further if the training data was cleaned prior to training. Also, using the sliding window technique and using an attenuating learning rate will improve the performance of the model.

Bzhalava et al [7] explored machine learning algorithms whilst using Relative Synonymous Codon Usage frequency (RSCU) for feature extraction. Random Forest and ANN were trained using two different datasets, which were used to classify the data samples into virus and non-virus groups. Two types of sequence databases were used: Codon Usage Database and a derived database. The second dataset was derived using Next Generation Sequencing (NGS) to generate metagenomic sequences from different patients. Random Forest and ANN were trained once with the data from the Codon Usage Database and secondly by the derived metagenomic dataset. The area under the Receiver Operating Characteristic (ROC) curve was used as the performance metric. To ensure a significant outcome was obtained, the leave-one-experiment-out cross-validation (LOEO) was implemented. Also, quality checking was performed prior to ensure it did not affect the analysis. They found that the machine learning techniques and RSCU is able to detect the viral contigs from the metagenomic data. However, the trained models using the derived data failed to generalize. Reasons for this are that noise was present in the dataset whereas the Codon Usage Database had clean data. Also, the performance is dependent on which features are used. Therefore, the use of more significant features may lead to a generalized model which can be used outside the known sequences. This will overcome the unclassified sequences giving rise to the misclassification loss due to their vast differentiation from the known sequences. They also found the codons that were responsible for the largest discrimination: TCG and CGC. However, it can be seen that there are limitations to these models and therefore suggests that further work should be focused on added flexibility by using pre-defined high-level features. This gives rise to the opportunity to find a novel feature that is able to classify viruses apart from its codon, which will help increase the classification accuracy.

Ren et al [8] in 2020, developed DeepVirFinder, a machine learning method, which identified viral sequences within an input metagenomic database. Their method outperformed the

previous method VirFinder [9]. They have used convolutional neural networks and taught them to learn viral genomic features and create a predictive model based on the learnt features. This method can be used for all contig lengths. The input metagenomic dataset used was RefSeq from the National Center for Biotechnology Information (NCBI). However, to increase the prediction accuracy further, the size of the training data was increased by adding unknown viral sequences from numerous other datasets. Later on, it was observed that this decision served another benefit of removing any sampling bias present and consequently increasing the prediction accuracy. Filtering of the samples was performed to ensure there was no contamination present that could affect the model's performance. Similarly, samples of high quality were selected only for training and validation to prevent any contamination affecting the model's performance. The performance of the model was evaluated using the area under the receiver operating characteristic curve (AUROC) metric. It was found that the DeepVirFinder had significantly enhanced prediction power such that it was able to predict viral sequences without the need of assembly. The model's robustness was tested further by checking its sensitivity to genetic mutations. They concluded that the model was not sensitive to ≤0.001 mutation rate. The DeepVirFinder is able to reliably identify viruses in metagenomic data. This methodology was applied to a case study to detect viruses in gut microbial communities for patients with CRC and the potential of DeepVirFinder was verified further as ten virus bins associated with cancer were detected.

Hsu et al [10] investigated how the LSTM model of deep learning can be implemented for the learning of genome patterns to be able to detect viruses from metagenomic data. They used a ground reference database as their input sequences. LSTM network (a type of recurrent neural network) was found to be suitable for the dealing with genomic sequences because the network is able to learn the long-term dependencies with respect to time in the input data. Therefore, after training, the network will have learnt the genomic patterns and therefore should be able to characterize unknown genomic sequences in the future. Hyperparameters were explored and implemented to increase the classification accuracy of the model. A further advantage, not previously available in the BLAST method, is that important segments of subsequences can be found without having the compare the complete input sequence. In addition, parallel computing concept was implemented to increase the computational speed of the model by executing processes concurrently. The performance of the model was evaluated in comparison with the BLAST method as well as in terms of accuracy and computational speed. They found that as the number of GPUs increased, the computational speed also increased linearly. The speed evaluated to be 36 times the speed of the BLAST method. Their overall results highlighted that similar accuracy was obtained to the BLAST method however the speed was significantly faster.

4 CONCLUSIONS

Over the last decade, a significant increase in the development of artificial intelligence techniques in the detection of viral genomes has been observed. Methodologies like VirNet and DeepVirFinder have overcome numerous limitations of previous tools and present advantages such as accuracy, speed, self-learning of features, whilst also providing further insights into the input metagenomic data. This presents an opportunity to investigate and explore further into the predictive ability of viral genomes for unknown viruses in an attempt to prevent the new virus from affecting the health of people.

REFERENCES

[1] Principles of Virology, 4th Edition, 2 Vol set by S. Jane Flint, Lynn W. Enquist, Vincent R. Racaniello, Glenn F. Rall, Anna Marie Skalka. .

[2] A. V.-L. S. C. January 06 and 2016, 'What Are Viruses?', livescience.com. https://www.livescience.com/53272-what-is-a-virus.html (accessed Oct. 31, 2020).

[3] T. Thomas, J. Gilbert, and F. Meyer, 'Metagenomics - a guide from sampling to data analysis', Microb Inform Exp, vol. 2, p. 3, Feb. 2012, doi: 10.1186/2042-5783-2-3.

[4] M. Margulies et al., 'Genome sequencing in microfabricated high-density picolitre reactors', Nature, vol. 437, no. 7057, pp. 376–380, Sep. 2005, doi: 10.1038/nature03959.

[5] G. Ditzler, R. Polikar, and G. Rosen, 'Multi-Layer and Recursive Neural Networks for Metagenomic Classification', IEEE Trans.on Nanobioscience, vol. 14, no. 6, pp. 608–616, Sep. 2015, doi: 10.1109/TNB.2015.2461219.

[6] A. O. Abdelkareem, M. I. Khalil, M. Elaraby, H. Abbas, and A. H. A. Elbehery, 'VirNet: Deep attention model for viral reads identification', in 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, Dec. 2018, pp. 623–626, doi: 10.1109/ICCES.2018.8639400.

[7] Z. Bzhalava, A. Tampuu, P. Bała, R. Vicente, and J. Dillner, 'Machine Learning for detection of viral sequences in human metagenomic datasets', BMC Bioinformatics, vol. 19, no. 1, p. 336, Dec. 2018, doi: 10.1186/s12859-018-2340-x.

[8] J. Ren et al., 'Identifying viruses from metagenomic data using deep learning', Quant Biol, vol. 8, no. 1, pp. 64–77, Mar. 2020, doi: 10.1007/s40484-019-0187-4.

[9] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, 'VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data', Microbiome, vol. 5, no. 1, p. 69, Jul. 2017, doi: 10.1186/s40168-017-0283-5.

[10] Y.-F. Hsu et al., 'High-Performance Virus Detection System by using Deep Learning', in 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, United Kingdom, Jul. 2020, pp. 1–9, doi: 10.1109/CEC48606.2020.9185808.