VirNet: A deep attention model for viral reads identification

Aly O. Abdelkareem
Department of Computer Engineering
Ain Shams University
Cairo, Egypt
aly.osama@eng.asu.edu.eg

Mahmoud I. Khalil
Department of Computer Engineering
Ain Shams University
Cairo, Egypt
mahmoud.khalil@eng.asu.edu.eg

Mostafa Elaraby
Microsoft Research
Cairo, Egypt
a-moelar@microsoft.com

Hazem M. Abbas
Department of Computer Engineering
Ain Shams University
Cairo, Egypt
hazem.abbas@eng.asu.edu.eg

Ali H. A. Elbehery
Institute of Virology
HelmholtzZentrum München
München, Germany
ali.elbehery@helmholtz-muenchen.de

***Abstract*—Metagenomics shows a promising understanding of function and diversity of the microbial communities due to the difficulty of studying microorganism with pure culture isolation. Moreover, the viral identification is considered one of the essential steps in studying microbial communities. Several studies show different methods to identify viruses in mixed metagenomic data using homology and statistical techniques. These techniques have many limitations due to viral genome diversity. In this work, we propose a deep attention model for viral identification of metagenomic data. For testing purpose, we generated fragments of viruses and bacteria from RefSeq genomes with different lengths to find the best hyperparameters for our model. Then, we simulated both microbiome and virome high throughput data from our test dataset with aim of validating our approach. We compared our tool to the state-of-the-art statistical tool for viral identification and found the performance of VirNet much better regarding accuracy on the same testing data.**

***Index Terms*—attention model, classification, deep neural networks, metagenomics, virus**

I. Introduction

Metagenomics is an analysis of the genetic information of the collective genomes of the microbes within a given environment based on its sampling regardless of cultivability of the cells. [8]. There is a minor population of microbial organisms identified due to the difficulty in studying them using pure culture isolation. This methodology has been constrained to less than 1% of host cells and is biased to certain species [10]. Metagenomic analysis process answers some questions about the identity of microorganisms in the collection and their potential functional characterization.

In this study, our interest is in prokaryotic microorganisms (e.g. bacteria and archaea) and viruses. Bacteria are unicellular and microscopic organisms that reproduce by binary fission. On the other hand, viruses are typically submicroscopic consists of genetic materials either DNA or RNA surrounded by a protective coat of proteins and can only replicate inside living host cells.

Viruses have an impact on different microbial communities . Phages or bacteriophages are viruses that infect bacteria. Furthermore, phages are abundant in different microbiome communities. Scientists are using isolation and culturing techniques to study viral diversity and viral-host interactions in microbial communities. Those techniques have many limitations because there is no universal marker gene for viruses. The sequenced viruses in NCBI RefSeq database constitute approximately 5% of known species of prokaryotic organisms [16].

High throughput sequencing technology is used for metagenomic studies which can generate large number of read sequences of microorganisms. Sequencing of microbial samples shows contamination of viral sequences within prokaryotic population. A study found 4-17% virus sequences in human gut prokaryotic metagenomes [12]. Moreover, cellular contamination is quite frequent even with a careful purification of viral particles.

The broadly adopted technique to know who is in metagenomic data is to assemble the high throughput reads to contigs then search against a known genomic database using sequence alignment method in order to infer the type of microorganisms and the existence of species in a metagenomic sample. This approach is minimal because it only detects viruses almost related to those we already know. It is reported that about 15% of viruses in the human gut microbiome and 10% in the ocean have similarity to the known viruses [13].

Machine learning approaches have been used to classify and cluster data based on extracted features. The deep neural network is one of machine learning methods that are considered as a state of the art category for general classification problems. Deep learning shows significant improvements in several artificial intelligence tasks for example image classification, speech recognition, and natural language processing. Additionally, It shows significant results with genomic data [2].

In this paper, we introduce a deep attention model, VirNet, to identify viral reads from a mixture of viral and bacterial sequences, and purify viral metagenomic data from bacterial contamination as well. That will guide us to identify new viruses and potentially perform functional characterization.

II. Related Work

There has been extensive prior work on viral identification. Recent work has focused on identifying phages in bacterial genomes. Several methods have used similarity search by sequence alignment with the reference genomes in order to find viral contigs. Most of the recent tools fall under three categories based on the sample structure such as:

1) phages from prokaryotic genomes
2) viral sequences in mixed metagenomic datasets
3) phages and viral sequences.

There are many software packages to find phages from prokaryotic genomes such as Phage_Finder [6], Prophinder [11], PHAST [18], and PhiSpy [1]. These tools are using similarity search to known virus databases using features such as genes or AT and GC skew.They have many limitations as they failed to detect viral sequences in metagenomic data as the databases are limited and don't represent viral diversity in the environment. Moreover, It is not optimized to process a large number of contigs [14] as they depend on sequence alignment processing limitations.

The second category is able to detect viral sequences in mixed metagenomic datasets such as VIROME [17] and MetaVir [15]. They are using similarity search with the databases same as the first category. Additionally, they are searching against proteins. There are more packages such as DIAMOND [4] or Centrifuge [9] which are much faster and efficient than the former tools for microbial classification.

The third category of software packages such as VirSorter [14] is able to detect phages and viral sequences. VirSorter is using similarity search to viral databases and integrates other features related to analysis of sequence genes such as enrichment of viral-like genes, enrichment of uncharacterized genes and viral hallmark gene. These features make the identification more accurate but it still suffer limitations. One of the limitations is the requirements of having at least 3 genes within the contig. Moreover, it is very slow in processing metagenomic datasets.

Recently VirFinder [13] applied machine learning techniques. VirFinder is a statistical method based on the logistic regression model. It uses the K-mer feature which is considered as a discrimination feature for different sequence problems. It shows a great success with short sequences too and They found a great k-mer similarity score with viruses within other prokaryotic genomes.

In this paper, we are using deep learning techniques which is more suitable to sequence problems and also shows significant improvements to other current machine learning models. In deep neural networks, the model will extract the most appropriate features during training which lead to better identification accuracy and sensitivity.

III. Materials and Methods

A. Building training and testing dataset

We divided viruses, bacteria and archaea genomes from RefSeq database randomly into a train and test genomes with 80% of total base pairs in training. Table I shows the number of genomes we used in training and testing. We processed all available viral genomes until Nov. 1st, 2017 and a sample from prokaryotic genomes due to the huge number of available prokaryotic genomes. Then, we converted the viral genomes into non-overlapping fragments of different lengths n = {100, 500, 1000, 3000}. We generated an approximate number of non-overlapping fragments of prokaryotic genomes with the same lengths randomly as well. (Table II).

Genome	Train	Test	Total
Viruses	7686	1870	9556
Prokaryotes	143241	35543	178784

TABLE I: The number of used genomes from RefSeq

Fragment Length (N)	Train	Test
100 bp	2088863	527020
500 bp	420857	106168
1000 bp	212253	53528
3000 bp	73163	18425

TABLE II: The number of fragments generated from viruses genomes

B. Generating simulated virome and microbiome

Grinder [3] is an open-source tool commonly used for generating a simulate amplicon and shotgun metagenomic datasets from reference genomes. We generated two metagenomic data of virome and microbiome of 1M reads and fragment length 100bp using Grinder with our reference test genomes to simulate shotgun metagenomic sequences in order to verify the ability of our tool to detect viral reads in metagenomic data instead of generated fragments from the reference genomes. The virome data has 75% of viral reads while microbiome has 25%.

C. Deep Learning Model

Recurrent neural networks (RNN), long short-term memory (LSTM) [7] and gated recurrent neural networks (GRU) [5] can model complex sequences and have been used for sequence modeling problems.

Our deep learning model is implemented as an attentional encoder network (Figure 1a). An input sequence $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_m})$ and calculates a forward sequence of hidden states $(\overrightarrow{h_1}, \ldots, \overrightarrow{h_m})$. The hidden states $\overrightarrow{h_j}$ are averaged to obtain the attention vector $\mathbf{h_j}$ representing the context vector from the input sequence.

Embedding layer maps discrete input words to dense vectors for computational efficiency before feeding this sequence to LSTM Layers. The attentional network could learn how to extract suitable features from the raw data and can attend to previous DNA nucleotide within the same input sequence.

The attentional neural model was trained with the DNA nucleotide bases with fragments with different lengths. The model will predict in a binary output format whether this fragment is viral or non-viral.

The top-performing model (Figure 1b) consists of an input embedding layer of size 128 mapping input DNA nucleotide tokens into an embedding space, that is fed to an LSTM layer. The forward sequence $\overrightarrow{h_j}$ is then averaged together to create an attentional vector representing token context within the same fragment. A dropout layer was added after the attentional layer to avoid overfitting over the input data.

The input sequence is divided into 5 grams sized tokens these tokens are then treated as a single word (Figure 1a). This single token is mapped as a point in the embedding space created during training the neural model.

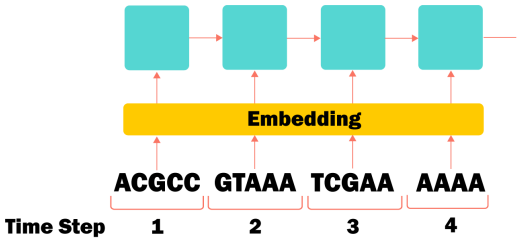
During training, all parameters are optimized jointly using Adam to maximize the conditional probability of tokens found together to predict if an input sequence is viral or not.

D. Hyperparameters optimization

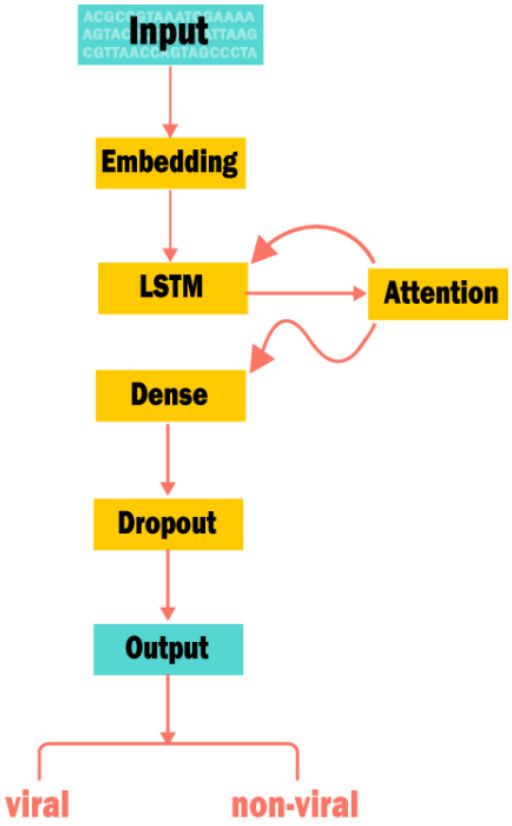
We selected the grid search technique in order to find the most suitable parameters. The grid search is considered a traditional technique for hyperparameters optimization and it brute force different combinations. We ran several experiments on 20% of our training set for 500 bp. Then, we divided it into training, validation, and testing set with the following percentages 70%, 10%, and 20%. These experiments were designed to find the best parameters for the number of recurrent layers, the embedding size for each layer and the input sub-words (ngram). Our reported results (Table III) show that the best parameters setup is for 2 layers, 128 embedding size, and 5 ngram.

#Layers	Embedding (#Neurons)	Ngram	ROC-AUC	Accuracy
1	32	3	0.8	73.66
1	32	5	0.83	76.42
1	32	7	0.79	72.11
1	64	3	0.83	76.1
1	64	5	0.83	75.98
1	64	7	0.77	69.96
1	128	3	0.83	75.76
1	128	5	0.85	77.41
1	128	7	0.78	70.93
2	32	3	0.8	73.25
2	32	5	0.83	76.49
2	32	7	0.79	72.82
2	64	3	0.81	73.96
2	64	5	0.84	76.46
2	64	7	0.78	72.53
2	128	3	0.83	76.15
2	**128**	**5**	**0.85**	**77.9**
2	128	7	0.78	70.63

TABLE III: Hyperparamters optimization results



(a) Embedding Layer



(b) Neural network model architecture

Fig. 1: VirNet model

IV. Results

A. Results for generated fragments

We tested VirNet on different lengths of fragments n= {100, 500, 1000, 3000} from our testing set of viruses and prokaryotes RefSeq genomes. Moreover, we compared the output results to VirFinder results on the same training and testing data. VirNet predictions outperformed VirFinder for fragments with length 500, 1000 and 3000 (Figure 2). The model reached to 82.82% of accuracy whereas VirFinder tool obtained 75.61%. Moreover, VirFinder can perdict the short fragments with length 100. Table IV shows the comparison between both tools in terms of accuracy.

Length(N)	Accuracy	
	VirNet	VirFinder
100	71.29%	63.9%
500	82.82%	75.61%
1000	86.82%	80.28%
3000	90.10%	87.11%

TABLE IV: Results comparison on fragments test-set

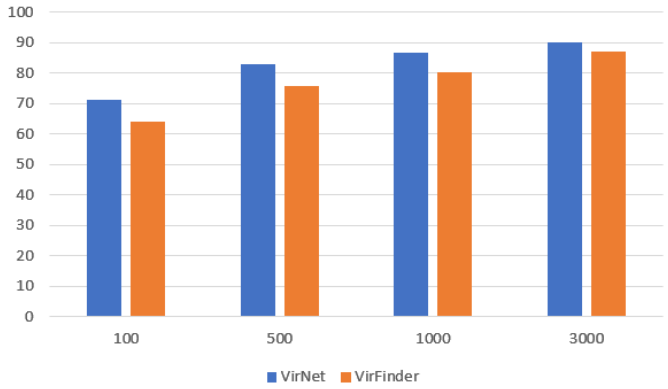


Fig. 2: VirNet vs VirFinder accuracy

B. Results for simulated metagenomes

As mentioned before, we tested VirNet on a simulated metagenomes of 100 bp and we found that VirNet performed much better than VirFinder. VirNet shows accuracy is 71.3% on the virome data and 72.14% on the microbiome data while VirFinder is 62.77% on the virome data and 64.49% on the microbiome data Table V).

Metagenome	Accuracy	
	VirNet	VirFinder
Virome	71.3%	62.77%
Microbiome	72.14%	64.49%

TABLE V: Results comparison on our simulated metagenomes

V. Discussion

In our tool, there are no handmade features as the network will learn how to extract appropriate features of the raw data. It shows better accuracy as it is trained with the updated viral databases with a good statistical model. This helps us to generalize this model with all genomes and to make a generalized model for sequence classification. There is no evidence that these training prokaryotic genomes don't have a viral infection or not. Cleaning the training genomes might give us better accuracy. For the trained deep learning model, using a sliding window over the input DNA sequence might improve our model, the only drawback of this technique is the slow training and inference time of input sequences, Also using an adaptive learning rate decaying over time steps during the training might improve the model performance, but will need more tuning over the input data.

VI. Conclusion

This attentional neural deep network was able to achieve state of the art results on viral identification from high throughput sequences. Our approach is able to classify short fragments as well. Experimental results validate our approach for identification with an accuracy 82.82% whereas VirFinder is 75.61% on 500 bp reads.

Availability of Data

VirNet is an open-source python package at https://github.com/alyosama/virnet. RefSeq genomes used are publicly available online via NCBI. All other generated data used in this study are available from the corresponding author on a request.

References

[1] Sajia Akhter, Ramy K Aziz, and Robert A Edwards. Phispy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic acids research*, 40(16):e126–e126, 2012.

[2] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[3] Florent E Angly, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic acids research*, 40(12):e94–e94, 2012.

[4] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59, 2014.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Derrick E Fouts. Phage_finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic acids research*, 34(20):5839–5851, 2006.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Jacques Izard and Maria Rivera. *Metagenomics for Microbiology*. Academic Press, 2014.

[9] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 2016.

[10] Jessica M Labonté, Brandon K Swan, Bonnie Poulos, Haiwei Luo, Sergey Koren, Steven J Hallam, Matthew B Sullivan, Tanja Woyke, K Eric Wommack, and Ramunas Stepanauskas. Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *The ISME journal*, 9(11):2386, 2015.

[11] Gipsi Lima-Mendez, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, 24(6):863–865, 2008.

[12] Samuel Minot, Rohini Sinha, Jun Chen, Hongzhe Li, Sue A Keilbaugh, Gary D Wu, James D Lewis, and Frederic D Bushman. The human gut virome: inter-individual variation and dynamic response to diet. *Genome research*, 2011.

[13] Jie Ren, Nathan A Ahlgren, Yang Young Lu, Jed A Fuhrman, and Fengzhu Sun. Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1):69, 2017.

[14] Simon Roux, Francois Enault, Bonnie L Hurwitz, and Matthew B Sullivan. Virsorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015.

[15] Simon Roux, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, Didier Debroas, and François Enault. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27(21):3074–3075, 2011.

[16] Simon Roux, Steven J Hallam, Tanja Woyke, and Matthew B Sullivan. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife*, 4:e08490, 2015.

[17] K Eric Wommack, Jaysheel Bhavsar, Shawn W Polson, Jing Chen, Michael Dumas, Sharath Srinivasiah, Megan Furman, Sanchita Jamindar, and Daniel J Nasko. Virome: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences*, 6(3):421, 2012.

[18] You Zhou, Yongjie Liang, Karlene H Lynch, Jonathan J Dennis, and David S Wishart. Phast: a fast phage search tool. *Nucleic acids research*, 39(suppl_2):W347–W352, 2011.