VirNet: A deep neural network for viral sequence identification

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

Mostafa Elaraby
Department of Computer Engineering
Alexandria University
Cairo, Egypt
mostafa.m.elaraby@gmail.com

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

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

Ali Elbehery
Institute of Virology
HelmholtzZentrum Mnchen
Mnchen, Germany
alielbehery@hotmail.com

***Abstract*—Metagenomics shows a promising understanding of function and diversity of the microbial communities due to the difficulty of studying microorganism with a pure culture isolation. Moreover, the viral identification is considered one of the most important steps in studying microbial communities. Several studies show different methods to identify viruses in mixed metagenomic data and phages in host genomes using homology and statistical techniques. These techniques have many limitations due to viral genome diversity. In this work, we propose a sequence deep model for viral identification of metagenomic data.**

Results: To test this method, we generate fragments of viruses and bacteria from RefSeq genomes with different lengths to find the best hyperparameters of our model. Then, We simulated both microbiome and virome high throughput data from our test-set genomes in order to validate our method. Finally, we applied our tool on a case study of two types of metagenomic data such as Roche 454 and Illumina. We found our sequence model reached 85.12% of accuracy whereas VirFinder tool obtained 75.61% with the same training and testing data.

Conclusion: We developed the first deep sequence network based on viral identification in large data. This tool will help us in expanding our knowledge in natural viral communities.

***Index Terms*—Metagenomics, Deep Learning, Viruses, LSTM, Classification**

I. Introduction

Metagenomics is the cultivation-independent analysis of the genetic information of the collective genomes of the microbes within a given environment based on its sampling [**?**]. There is a small fraction of extant microbial life has been identified because of the difficulty to study microbiology with a pure culture isolation. Scientist reported that there are less than 5% of microorganisms can be easily cultured(Need a reference). Metagenomic analysis process demonstrates a promising understanding of different microorganisms. It answers some questions about microorganisms in the collection such as Who is there?, What can they do? and what can they potentially do?.

Microorganisms are found everywhere on earth and they are very important in our life. They are essential to our life. 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 and virus-host interaction can change many ecosystems such as human health and aquatic life. Phages are viruses that mainly infect bacteria and archaea. Furthermore, phages are abundant in Microbiome. Scientists are using isolation and culture-independent techniques to study viral diversity and viral-host interactions in microbial communities. Those techniques have many limitations because of there is no universal marker gene for viruses. Some reported most of the sequenced viruses in NCBI RefSeq are from 5% of known phyla of prokaryotic hosts [**?**].

High throughput sequencing technology is used for metagenomic studies which can generate massive amounts of short read sequences from prokaryotic cells in microbial communities regardless of cultivability of the cells, and viruses are inevitably captured at the same time in these samples. Sequencing microbial samples found viral sequences along with prokaryotic hosts. A study found 4-17% virus sequences in human gut prokaryotic metagenomes [**?**]. Moreover, Cellular contamination is quite frequent even with a careful purification of viral particles and this is one of the main reasons why we need a tool that can differentiate between bacterial and viral sequences.

The classical approach to know who is in metagenomic data is to assemble the high throughput reads to contigs then search against know genomic database in order to know these microorganisms. This approach is very limited because it only finds viruses closely related to those we already know about. It is estimated that only about 15% of viruses in the human gut microbiome and 10% in the ocean have similarity to the known viruses [?].

Machine learning approaches have been used to classify and cluster data based on hand-made 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. Moreover, It shows promising results with genomic data [?].

In this paper, we introduce a deep sequence model, VirNet, to identify viral sequences from a mixture of virus and bacterial sequences and purify viral metagenomic data from bacterial contamination as well. This will lead to discover more new viruses and help us to understand their functionality and diversity in the ecosystem.

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 with the known genomes in order to find viral contigs. There are three types of recent tools:

1) Phages from prokaryotic genomes
2) Viral sequences in mixed metagenomic datasets
3) Phages and Viral sequences.

There are many methods to find phages from prokaryotic genomes such as Phage_Finder [?], Prophinder [?], PHAST [?], and PhiSpy [?]. These tools are using similarity search to known viruses databases using features such as genes. Some of them such as PhiSpy integrates other features such as protein length, AT and GC skew, transcription strand direction and unique virus k-mers. They have many limitations as they failed to detect viral sequences in metagenomic data as the databases are outdated, limited and don't represent viral diversity in the environment. Moreover, It is not optimized to process a large number of contigs [?] as they depend on alignment and homology processing limitations.

The second type is able to detect viral sequences in mixed metagenomic datasets such as VIROME [?] and MetaVir [?]. They are using similarity search with the databases as same as the first type and Also, they are searching against proteins. Some people are using DIAMOND [?] or Centrifuge [?] as they are fast and efficient for microbial classification. The limitation of this approach is related to limited databases as they only search for known genomes.

The third type is able to detect phages and viral sequences such as VirSorter [?]. VirSorter is using similarity search to viral databases and integrates other features such as viral hallmark gene, enrichment of viral-like genes, enrichment of uncharacterized genes which make it more accurate but it has many limitations such as it require at least 3 genes within the contig as the smallest virus genome contains 3 genes so it has the same limitations as previous techniques because of using homology strategy. Moreover, it cannot work with short fragments or contigs and it is very slow in processing metagenomic datasets.

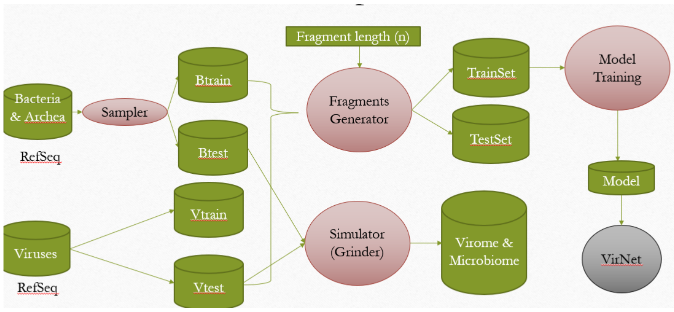


Fig. 1. VirNet Data Pipeline

Recently VirFinder [?] 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 much more suitable to sequence problems and also shows significant improvements to other machine learning models. In deep neural networks, the model will extract the most appropriate features during training which lead to better identification accuracy.

III. Materials and Methods

We downloaded Viruses, Bacteria and Archaea genomes from RefSeq database then we split it into a train, valid and test genomes. Then, We took random non-overlapping fragments of different lengths (n= 100,150,300,500,1000). We divided the data into viral and non-viral with an approximate number to make the data balanced as the DNN models don't learn well from unbalanced data. Figure 1 shows the data pipeline. Table I is the number of genomes we used in training and testing. We downloaded all the viruses genomes until Nov. 1st, 2017 but we sampled bacterial genomes.

	Viruses	Bacteria & Archaea
Total	9556	178784
Train	7686	143241
Test	1870	35543

TABLE I
RefSeq Genomes

Moreover, We generated two metagenomic data Virome and Microbiome of 1M reads (100bp) using Grinder [?] with our test genomes to simulate shotgun metagenomic sequences. Additionally, We used Illumina error model indicated by mutation_dist poly4 3e-3 3.3e-8 and mutation ratio 91:9 (9 indels for each 91 substitution mutations) because for Illumina indel errors occur more often than substitution errors

[?]. Table II shows simulated data statistics.

	Microbiome	Virome
Bacteria Length	75450367 bp	17551396 bp
Bacteria Genomes	1488	422
Bacterial reads	803742	176059
Viruses Length	25133078 bp	52609236 bp
Viruses Genomes	845	1870
Viral reads	196258	823941
Viral Ratio	25.00%	75.00%
Library coverage	0.994x	1.001x
Diversity (richness)	2302	2726

TABLE II
GRINDER SIMULATED METAGENOME

Then, we applied our tool to two real metagenomes as a case study

1) **454**: Subtropical freshwater microbial and viral metagenome (SRR648314).
2) **Illumina**: Lake Michigan virome (SRX995836).

Our tool is able to read not only fasta files and fastq files. Furthermore, it is able to deal with paired-end reads i.e. if one of the two pairs is identified as a virus, the other should be the same. If there are conflicts between the classifications of the two pairs; this pair could be denoted as ambiguous.

IV. DEEP LEARNING MODEL

The Deep Learning system is implemented as an attentional encoder network 2. 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/GRU 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.

Recurrent neural networks (RNN), long short-term memory (LSTM) [?] and gated recurrent neural networks (GRU) [?] can model complex sequences and have been used for sequence modeling problems. All these variations were used during to get the performing model over the input fragments.

LSTM encoder have 3 gates to protect and control the cell state as in 3,the input gate denoted $\mathbf{i}$ which defines how much of the newly computed state you want to let through, forget gate denoted $\mathbf{f}$ that decides what information is to be kept and what is to be thrown away , the output of the update gate denoted $\mathbf{U}$ that's used to update the cell state and the output of the LSTM cell $\mathbf{o}$.$\mathbf{W}$ is the recurrent connection at the previous hidden layer and current hidden layer and $\mathbf{C}$ is the internal memory of the unit as shown in the following equations

$$\mathbf{i_t} = \sigma(\mathbf{x_t U^i} + \mathbf{h_{t-1} W^i})$$
$$\mathbf{f_t} = \sigma(\mathbf{x_t U^f} + \mathbf{h_{t-1} W^f})$$
$$\mathbf{o_t} = \sigma(\mathbf{x_t U^o} + \mathbf{h_{t-1} W^o}).$$

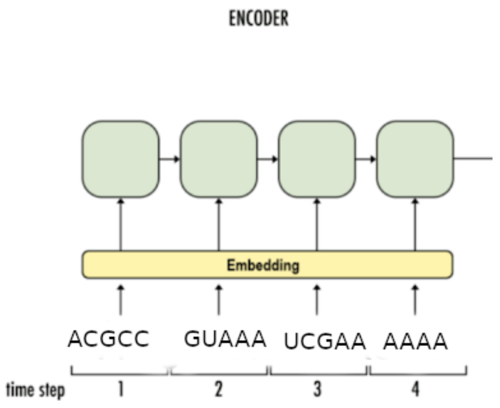


Fig. 2. Embedding Layer

GRU encoder 3 is same as LSTM except it has only 2 gates, Reset gate denoted $\mathbf{r}$ that determines how to combine the new input with the previously saved input state and the update gate denoted $\mathbf{z}$ that defines the amount of information to keep around, as defined in the following equations

$$\mathbf{z_t} = \sigma(\mathbf{x_t U^z} + \mathbf{h_{t-1} W^z})$$
$$\mathbf{r_t} = \sigma(\mathbf{x_t U^r} + \mathbf{h_{t-1} W^r})$$
$$\overline{\mathbf{h_t}} = \mathbf{tanh}(\mathbf{x_t U^h} + (\mathbf{r_t} * \mathbf{h_{t-1}})\mathbf{W^h})$$
$$\mathbf{h_t} = (\mathbf{1} - \mathbf{z_t}) * \mathbf{h_{t-1}} + \mathbf{z_t} * \overline{\mathbf{h_t}}.$$

The few number of gates in GRU makes it faster than the LSTM cell and based on the data might have better performance. In our model after validating various type of models the GRU attentional model performed well on the input DNA nucleotide.

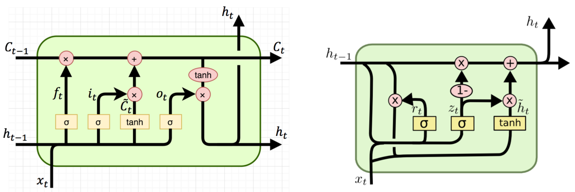
The attentional neural model 4 was trained with the DNA nucleotide bases with 100bp fragments.The model will then try to predict in a binary format whether this fragment is viral or non-viral.

The top performing model consists of an input embedding layer of size 128 mapping input DNA nucleotide tokens into an embedding space , that's fed to a GRU layer.The forward sequence $\overrightarrow{h_j}$ is then averaged togethor 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 single word .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.

In this model , an early stopping mechanism was used as a form of regularization to avoid overfitting over the data while making more epochs over the data. The early stopping mechanism was used with patience of 3 non improving consecutive epochs , the neural model will stop training while saving the latest improving checkpoint over the validation set defined.



(a) Long Short-Term Memory

(b) Gated Recurrent Unit

Fig. 3. LSTM and GRU encoders

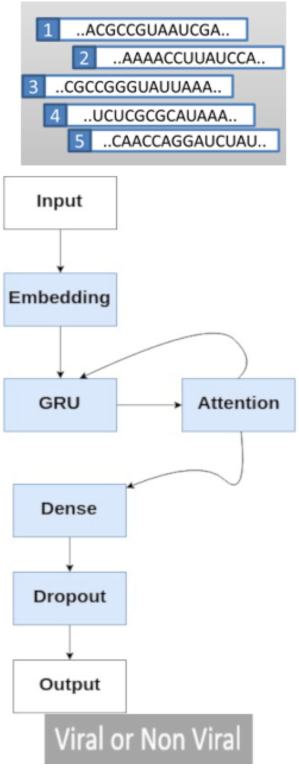


Fig. 4. VirNet Model architecture

V. Results

To test this method, we generate fragments of viruses and bacteria from RefSeq genomes with different lengths to find the best hyperparameters of our model. Then, We simulated both microbiome and virome high throughput data from our test-set genomes in order to validate our method. Finally, we applied our tool on a case study of two types of metagenomic data such as Roche 454 and Illumina. We found our sequence model reached 85.12% of accuracy whereas VirFinder tool obtained 75.61% with the same training and testing data. Table III shows experiment results.

We tried different fragment length (n=100,500,1000,3000) with VirFinder and we found that VirFinder cannot identify viruses 100 bp. See figure 5 and table IV.

We tried VirFinder on the simulated metagenomes and we found that VirFinder accuracy is arround 65%. See figure 6 and table V.

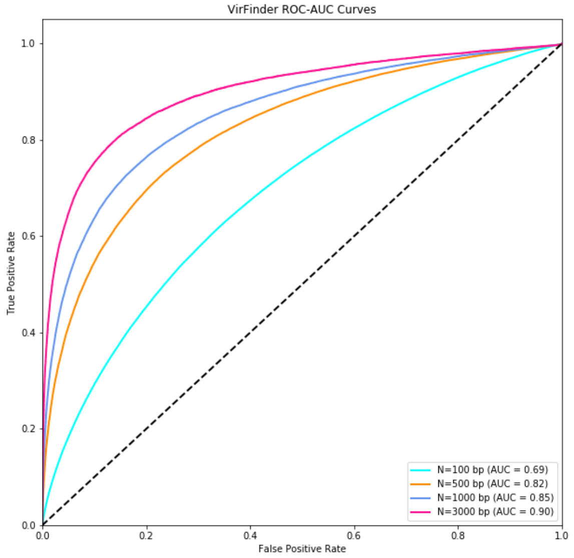


Fig. 5. VirFinder ROC-AUC curves

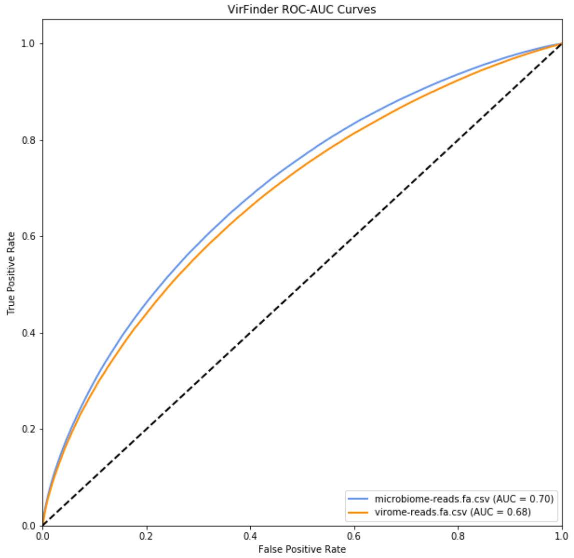


Fig. 6. VirFinder ROC-AUC curves on Simulated metagenomes

Model	Accuracy	ROC-AUC
LSTM (1L , 128 N) + DNN (64 N)	79.80%	80%
LSTM (2L , 128 N) + DNN (64 N)	82.88%	83%
LSTM (3L , 128 N)	80.32%	81%
LSTM (2L, 64 N)	79.20%	79.50%
LSTM (2L, 256 N)	78.70%	79%
RNN (2L)	81.30%	81%
GRU (2L)	82.10%	82%
LSTM (2L, 128 N, 0.2 dropout)	**85.12%**	85%
VirFinder (Benchmark)	**74.40%**	74%

TABLE III
HYPERPARAMTERS OPTIMIZATION RESULTS

Length(N)	Accuracy	Avg. Precision	Avg. Recall	ROC-AUC Score
100	63.9%	0.64	0.64	0.64
500	75.61%	0.76	0.76	0.75
1000	80.28%	0.82	0.80	0.78
3000	87.11%	0.88	0.87	0.83

TABLE IV
VIRFINDER RESULTS ON OUR TEST-SET

Length(N)	Accuracy	Avg. Precision	Avg. Recall	ROC-AUC Score
Virome	62.77%	0.72	0.63	0.63
Microbiome	64.49%	0.72	0.64	0.64

TABLE V
VIRFINDER RESULTS ON OUR SIMULATED METAGENOMES

VI. DISCUSSION

In our tool, there are no handmade features as the network will learn how to extract appropriate features for 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. We also able to identify the short viral sequences as LSTM learns from the dependences between the input sequence. There is no evidence that these training prokaryotic genomes don't have viral infection or not. We need clean the training genomes to get better accuracy. Additionally, Using GPUs in this tool make it very fast and scalable with a large number of sequences. We found VirNet 16X faster than normal methods.

VII. CONCLUSION

We developed the first deep sequence network based on viral identification in large data. This tool will help us in expanding our knowledge in natural viral communities.

ACKNOWLEDGMENTS

We would like to thank everyone for their help and supervision.