VirNet: Sequence model for viral identification

***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 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 neural 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 of our model. Then, we simulated both microbiome and virome high throughput data from our test genomes dataset with aim of doing validate our approach. We compared our tool to the state-of-the-art statistical and popular tool for viral identification and found the performance of VirNet much better regarding accuracy and speed on the same testing data. This tool will help us in growing our insights into natural viruses of microbial communities.**

***Index Terms*—metagenomics; deep learning; virus; sequence model; classification**

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. [9]. There is a little portion 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 [11]. 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? Besides, what can they potentially do?.

Microorganisms are found everywhere on earth, and they are critical in our life. They are essential to our survival. 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 as they don't have metabolic enzymes and equipment such as ribosomes for making proteins. There are 200 to a few thousand genes in the bacterial genomes, while the tiniest viral genomes have only three genes and the largest have hundred to 2000 genes.

Viruses have an impact on different microbial communities, and virus-host interaction can change many ecosystems such as human health and aquatic life. Phages or bacteriophages are viruses that infect bacteria. Furthermore, phages are abundant in different microbiome communities. The viral infection starts when virus binds to a host cell and its genome integrated with the host cell genome. The integrated viral DNA called a provirus. It is reasonable to find that isolated viruses are just package of genes want to move from one host cell to another. 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 there is no universal marker gene for viruses. Some reported most of the sequenced viruses in NCBI RefSeq are from 5% of known species of prokaryotic organisms [18].

High throughput sequencing technology is used for metagenomic studies which can generate large number of read sequences of microorganisms. The expected read length is up to 600 bp and the number of generated reads per run is up to 15 million approximately based on the sequencer platform and used kit [2]. We can sequence prokaryotic and viral cells in microbial communities in a cultivation-independent process 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 [14]. 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 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 [15].

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. Moreover, It shows significant results with genomic data [3].

In this paper, we introduce a deep sequence model, VirNet, to identify viral sequences from a mixture of viral and bacterial sequences and purify viral metagenomic data from bacterial contamination as well. That will guide us to find new viruses and study their genomes, and beside of that, it answers many mysteries related to our understanding of 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 [7], Prophinder [13], PHAST [20], and PhiSpy [1]. 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 unique virus k-mers, AT and GC skew, protein length and transcription strand direction. 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 [16] 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 [19] and MetaVir [17]. 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 [5] or Centrifuge [10] 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 [16]. 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 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 because the smallest virus discovered has 3 genes only 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.

Recently VirFinder [15] 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

A. Building training and testing dataset

We downloaded viruses, bacteria and archaea genomes from RefSeq database then we divided them 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 used 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). We balanced the data of both classes using random under-sampling technique to avoid the deep neural network bias to the majority class. Figure 1 shows more details for data pipeline.

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 [4] 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 work on metagenomic data instead of fragments from the genomes. Moreover, 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 [12]. Table III shows simulated data statistics.

C. Deep Learning Model

The deep learning model is implemented as an attentional encoder network (Figure 2a). 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.

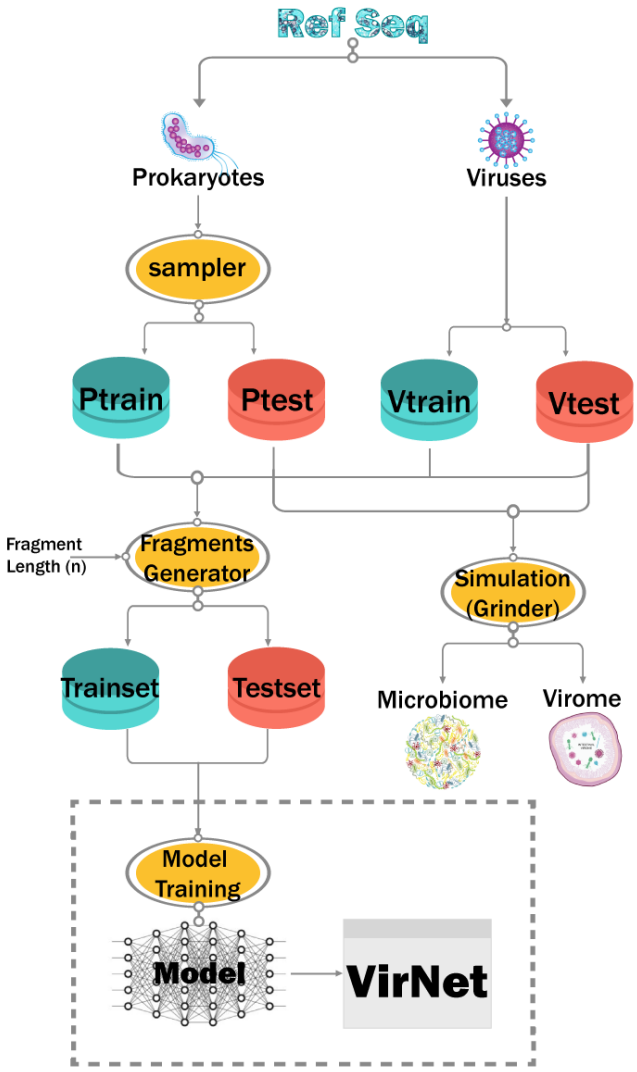


Fig. 1: VirNet Data Pipeline

	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 III: Grinder Simulated Metagenome

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) [8] and gated recurrent neural networks (GRU) [6] can model complex sequences and have been used for sequence modeling problems.

LSTM encoder have 3 gates to protect and control the cell state, 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

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

GRU encoder 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 attentional neural model 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 (Figure 2b) 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.

LSTM layer was performing better as in IV than the GRU cell.GRU encoder having less gates than LSTM model make it faster and easier to converge , but depedending on the size and the format of the input data LSTM with more gates would be slower but will outperform GRU encoder type.

The input sequence is divided into 5 grams sized tokens these tokens are then treated as a 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 over-fitting 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.

D. Hyperparameters optimization

Model parameters affect the performance of the deep learning model, and they control the behavior of the training algorithm as well. 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 IV) shows that the best parameters setup was for 2 layers, 128 embedding size, and 5 ngram.

We ran other experiments with the same parameters to check the ability of our model with other configuration, so we changed different parameters separately such as embedding size to 256 neurons, the number of layers to 3 and the recurrent cell type to GRU instead of LSTM. We found a slightly less accuracy and roc-auc scores but the training time was much more than the best parameters configuration as expected due to increasing number of neural network parameters.

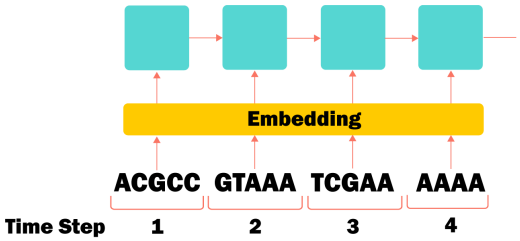
#Layers	Embedding (#Neurons)	Ngram	ROC-AUC	Accuracy
1	32	3	0.8	73.66
1	32	5	0.83	76.42
1	32	7	0.79	72.11
1	64	3	0.83	76.1
1	64	5	0.83	75.98
1	64	7	0.77	69.96
1	128	3	0.83	75.76
1	128	5	0.85	77.41
1	128	7	0.78	70.93
2	32	3	0.8	73.25
2	32	5	0.83	76.49
2	32	7	0.79	72.82
2	64	3	0.81	73.96
2	64	5	0.84	76.46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

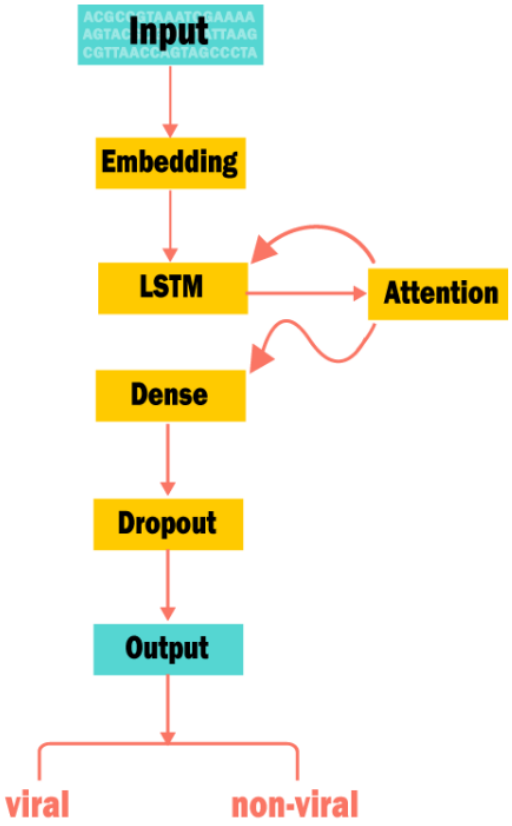
IV. Results

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 3). The model reached to 82.82% of accuracy whereas VirFinder tool obtained 75.61%. Moreover, VirFinder can perdict the short fragments with length 100. Figures 4a and 4b shows roc-auc curves of both tools on the testing set. Table V shows the comparison between both tools in terms of accuracy, precision, recall and roc-auc scores.



(a) Embedding Layer



(b) Neural network model architecture

Fig. 2: VirNet model

Length(N)	Accuracy		Avg. Precision		Avg. Recall	
	VirNet	VirFinder	VirNet	VirFinder	VirNet	VirFinder
100	71.29%	63.9%	0.72	0.64	0.71	0.64
500	82.82%	75.61%	0.83	0.76	0.83	0.76
1000	86.82%	80.28%	0.87	0.82	0.87	0.80
3000	90.10%	87.11%	0.91	0.88	0.90	0.87

TABLE V: VirFinder results on fragments test-set

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 70% while VirFinder is 62.77% on the virome data and 64.49% on

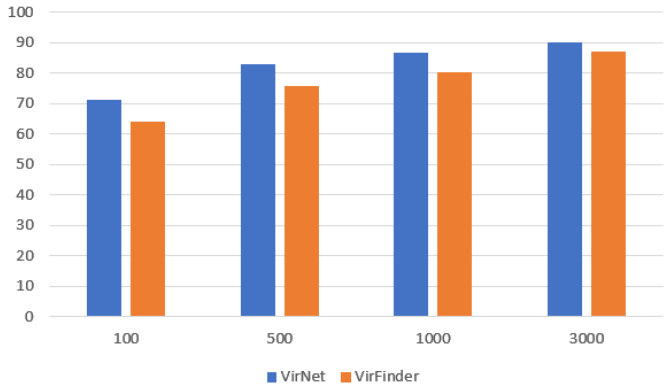


Fig. 3: VirNet vs VirFinder Accuracy

the microbiome data (Figure 4c and Table VI).

Metagenome	Accuracy		Avg. Precision		Avg. Recall	
	VirNet	VirFinder	VirNet	VirFinder	VirNet	VirFinder
Virome	71.3%	62.77%	0.71	0.63	0.72	0.62
Microbiome	72.14%	64.49%	0.73	0.65	0.73	0.64

TABLE VI: VirFinder results on our Simulated Metagenomes

C. VirNet processing speed

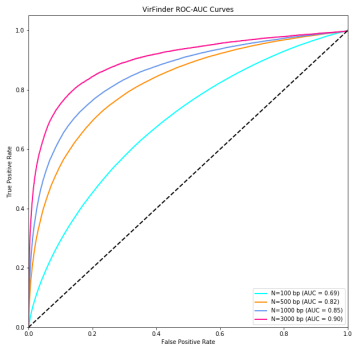
Using GPUs is an advantage for our tool and make it very fast and scalable with a large number of reads and fragments. The training process is around 5 hours with 2 million reads and the prediction is 30 seconds on 1 million reads using Nvidia GeForce GTX 1080 Ti. On the other hand, VirFinder processing the reads on a single CPU less than VirNet by around 82 times. We can avoid that by using parallel CPU threads for VirFinder but in case you want to retrain it with new data, it will take a few days.

V. Discussion

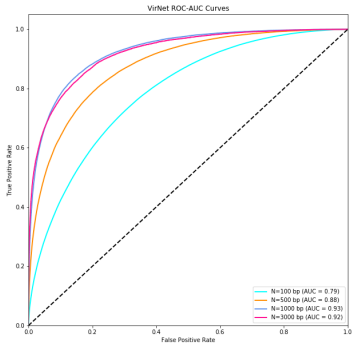
In our tool, there are no handmade features as the network will learn how to extract appropriate features for the raw data. It shows better accuracy as it is trained with the updated viral databases with a good statistical model. This helps us to generalize this model with all genomes and to make a generalized 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 a viral infection or not. Cleaning the training genomes would give us better accuracy but based on sampling and randomizing prokaryotic fragments, our training data might not contain proviruses.

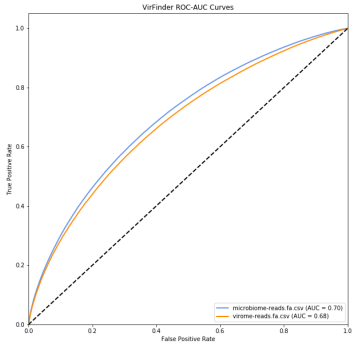
For the trained deep learning model, using a sliding window over the input DNA sequence might improve our model , the only drawback of this technique is the slow train and inference time of input sequences, Also using an adaptive learning rate decaying over time steps during the training might improve the model performance, but will need more tuning over the input data.



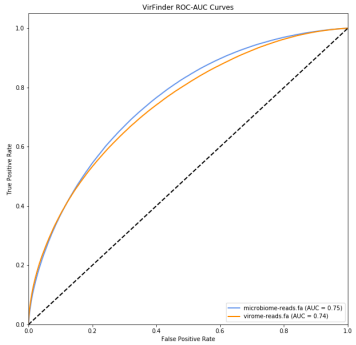
(a) VirFinder performance on generated fragments



(b) VirNet performance on generated fragments



(c) VirFinder performance on the simulated genomes



(d) VirNet performance on the simulated genomes

Fig. 4: VirFinder ROC-AUC curves on fragments and simu-

VI. Conclusion

This attentional neural deep learning 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 of more than 83%. According to these results, Our model would help us in understanding viruses in various microbial communities and discovering new species of viruses.

Acknowledgments

We acknowledge Eng. Mayada Elsayed for designing the figures and Dr. Kiran Natra for the support.

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] Imane Allali, Jason W Arnold, Jeffrey Roach, Maria Belen Cadenas, Natasha Butz, Hosni M Hassan, Matthew Koci, Anne Ballou, Mary Mendoza, Rizwana Ali, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC microbiology*, 17(1):194, 2017.

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

[4] 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.

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

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] David Laehnemann, Arndt Borkhardt, and Alice Carolyn McHardy. Denoising dna deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in bioinformatics*, 17(1):154–179, 2015.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.