Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023429

Using sequence-to-sequence LSTM to predict RNA virus mutations

Lianghong Chen

Computer science, Western University, 1151 Richmond St, London, ON N6A 3K7, Canada

lchen776@uwo.ca

Abstract. Infection with viruses is one of the main causes of human illness and even death in today's society. Vaccination can help people fight the virus. However, virus mutation will always cause vaccines to fail. Predicting the mutations of viral could help detect the mutated virus early and develop new vaccines so to reduce the rate of infection and death from the virus. The common strategy used to predict virus mutation is determining the important components of the virus, like amino acids and proteins, then using deep learning or machine learning methods to construct models. It is an effective strategy. However, using this strategy always cause people to spend lots of time and money studying the component of the virus, like amino acids, proteins, nucleic acid and so on. To increase efficiency and reduce the cost of research, a new method that can predict virus mutation based on nucleotide sequence alone is something people looking forward to. In recent years, in natural language processing, building a sequence-to-sequence model is becoming a popular and effective method to deal with textual data. As a type of text data, using the idea of sequence to sequence should be good to deal with the RNA sequence. In addition, in terms of the past study of predicting virus mutation, RNA sequence as a kind of time sequence, people generally use long short-term memory (LSTM) methods to handle it. Thus, in this study, we would combine the idea of the sequence-to-sequence model with LSTM method to predict the possible mutation of the virus. This experiment studies the mutation of two typical influenza viruses and achieves encouraging results.

Keywords: RNA virus, mutation, sequence-to-sequence LSTM.

1. Introduction

1.1. Virus

As it is estimated that there are more than 35 million people inflect with human immunodeficiency virus (HIV) and millions of people die of HIV [1,2]. HIV is a kind of RNA virus. RNA virus like HIV has a strong mutational potential, which means it generally has lots of subtypes. This is one of the main reasons that people now cannot cure Acquired Immune Deficiency Syndrome. If people can find a way to predict virus mutations accurately, it will not only help people to learn more about HIV but also be helpful to learn about other viruses using similar ways. If people can learn about viruses enough, it is more possible to develop effective ways to prevent virus infection or cure infected people, like developing vaccines [2-4]. Except for RNA viruses like HIV mentioned above, there is another kind of

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/4/2023429

virus DNA virus, which has DNA as genetic material with a double-stranded structure. DNA viruses include hepatitis B virus, smallpox virus, and so on. Compared with DNA viruses, RNA virus is not stable, since the genetic material of RNA virus is single-stranded rather than double-stranded structure. Generally, the virus with single-stranded structural genetic material is less stable than the virus with double-stranded structural genetic material. However, the more unstable the virus, the more likely it is to mutate. Therefore, RNA viruses are easier to mutate than DNA viruses [1]. If people want to study the mutation of viruses, the RNA virus is more worthy to study than the DNA virus.

1.2. Mutation

A virus mutation is a change in the genetic material of a virus, which always produces a new sub-type virus. The mutated virus might disable the vaccine for the virus since a vaccine can only be used to prevent a specific virus generally not including its sub-types. Even if the vaccine can prevent some sub-types of the virus, its effectiveness may be diminished. In other words, the mutated virus may infect humans although these people have got the vaccine for the original virus [5-8]. Thus, if people can develop vaccines for mutated viruses before the virus actually mutate, then people can get these new vaccines early and prevent virus inflection more effectively. To develop new vaccines for the mutated virus, people should know the basic information about the mutated virus first, especially its nucleotide sequence. So, the target of this experiment is to predict the possible nucleotide sequence of the mutated virus to provide a reference for developing vaccines. To reach this target, the experiment uses a deep learning method called sequence-to-sequence LSTM to predict the possible nucleotide sequences of mutated viruses.

1.3. Sequence-to-sequence long short-term memory methods

Sequence-to-sequence long short-term memory model is a kind of special sequence-to-sequence model, which adds LSTM layers to the traditional sequence-to-sequence model.

Sequence-to-sequence model is a kind of deep learning model, which can be applied to machine translation, text summarization, video captioning and so on. Sequence-to-sequence models can process the data in a sequence structure and the component of the sequence could be words, numbers, letters and so on. The output of the sequence-to-sequence model is also in a sequence structure. For example, in machine translation, inputting a sentence which is a word sequence and then the translation result from the trained model will be a sentence consisting of a sequence of words as well. The sequence-to-sequence model has two core parts an encoder and then following a decoder. The encoder recognizes the context of the input data and constructs the hidden state vector for the input. After the conversion is done, the encoder sends the hidden state vector to the decoder. After the decoder receives the input, the decoder can then compute and produce the output. Since the input of the task is a sequence structure data, the encoder and decoder generally should be made of methods which can deal with sequence input like RNN, LSTM, GTU and so on [9, 10].

The rationale of LSTM is very similar to the recurrent neural network (RNN), which makes the problems like gradient explosion and gradient no longer arise in the progress of training complex and long sequences. This indicates that LSTM has better performance than the normal RNN model when the trained sequence has complex long-term dependencies. The key to LSTM is states, including hidden states and cell states. The hidden states in LSTM are like hidden states in RNN, which represent short-term memory and cell states can be used to represent long-term memory. The core of LSTM consists of three parts, including the forget gate for remembering or forgetting information and the input and output gate for delivering information. All these three gates are used to operate states in LSTM. The forget gate in LSTM is used to determine whether keep the information with the previous timestamp. To be specific, the formula of the forget gate will compute a value as a criterion. If the computed value is 0, the network will remember nothing and if computed the value is 1, the network will remember everything. The function of the input gate is to determine the importance of the carried information. The formula in the input gate will compute a value between -1 and 1. If the value is a negative value, the carried information will be removed from cell states. Otherwise, it will be added to the cell states. The function of the output

DOI: 10.54254/2755-2721/4/2023429

gate is computing the predicted result. In this gate, the current hidden state will be computed according to the current cell state. Then, the Softmax function is going to activate the hidden state. The main step of this gate is using cell states to update the hidden and use the Softmax function to compute the result [11].

Sequence-to-sequence long short-term memory model combines the sequence-to-sequence model with the long short-term memory method together. Specifically, this model has an encoder and a decoder both being made of LSTM layers. The LSTM layers in the encoder summarize the input information and output the cell states and hidden states. The decoder will receive the outputs from the encoder and produce the next-generation sequence according to the outputs. The next-generation sequence is the predicted result looking forward to [12, 13].

2. Related work

Predicting the mutation of the virus is not a new topic. Using artificial intelligence methods to predict virus mutations has been practiced in genetics for a long time. A common strategy to predict virus mutations by using artificial intelligence methods is to construct models according to the relationships between the cause of the mutation and the mutation site. To be specific, scientists may select some features like unique amino acids and proteins in the virus and take mutation sites as targets, using some machine learning or deep learning methods to find the stable relationship between features and mutation sites [14]. Here are two typical studies.

Deep learning is often used to predict virus mutations [15]. Xia et al. predicted the H3N2 virus mutation by modelling the amino acid sequences of influenza viruses. Amino acids are the result of the expression of viral nucleotide sequences. The accumulation of mutations can be recognized by changes in amino acids in the virus, which can be used to predict the mutation of the virus. Xia's team uses the integrated model consisting of the convolutional neural networks (CNN) and the bidirectional LSTM model to predict virus mutation. They use CNN to get the complex local amino acid sequences and BLSTM to capture all the rest of the amino acid sequences. Their method performs high accuracy on the data sets used for validation. In addition, high accuracy was achieved when the model was used to predict mutated viruses in the next one or two years. Thus, deep learning methods can be used to predict some virus mutations [16].

Some machine learning methods like random forests and logistic regression can also be used to predict mutations of some influenza viruses [17, 18]. Yao et al. compared 95 alternative matrices of amino acid properties correlating with influenza virus antigenicity predicted by random forest models. Then, they modify the traditional random forest regression so that the model can jointly deal with the replacement matrix at the top. The RNA sequence used in the experiment is the H3N2 human seasonal influenza virus, which is from the period between 1968 and 2003. They use this algorithm to predict the possible mutations of the virus. By using some validation strategies like 10-fold cross-validation, they prove this algorithm is better than other common methods in predicting the mutation of the virus antigenic. Moreover, their experiment also shows that the structural features are the major factor causing the mutation of correlating influenza antigenic. After analyzing data from two adjacent antigenic clusters, they deduced several key amino acids causing mutations. Finally, Yao constructs an antigenic map for all H3N2 viruses researched, which can help to identify the related pathogens quickly [14].

The above two types of methods demonstrate the great potential of deep learning and machine learning in predicting RNA virus mutations. Nonetheless, it is worth noting that their common ground is needing to extract amino acid or protein features that may cause viral mutations and then construct models to find the relationship between these features and mutations. Although these experiments achieve high accuracy and have certain interpretability, they rely on multivariate data and need a long time to research. To get amino acids, proteins and other components data of the virus, people need to spend more time and more scientific research funds compared with only extracting the nucleotide sequence of the virus. However, if the mutation can be predicted based on its nucleotide sequence alone in the future, predicting virus mutations would be easy, efficient, cheap and no longer rely on the data

DOI: 10.54254/2755-2721/4/2023429

such as amino acids. Therefore, this experiment will try to predict the possible virus mutations according to the virus' original nucleotide sequence and expect to achieve high accuracy.

3. Method

3.1. The dataset

The viruses studied in this study are two typical influenza viruses, H3N2 and H1N1. The datasets are some RNA sequences of the viruses, and they are from the Medline database. The Figure 1 and Figure 2 are the samples of the RNA sequences.

```
caaaaacttc ccggaaatga caacagcacg gcaacgctat gccttgggca ccatgcagta
ccaaacggaa cgatagtgaa aacaatcaca aatgaccaaa ttgaagttac taatgctact
gagctggttc agagttcctc aacaggtgga atatgcgaca gtcctcatca gatccttgat
ggagaaaact gcacactaat agatgctcta ttgggagacc ctcagtgtga tgacttccaa
aataagaaat gggacctttt tgttgaacgc agcaaagcct acagcaactg ttacccttat
gacgtgccgg attatgcctc ccttaggtca ctagttgcct catccggcac actggagttt
aacaatgaaa gcttcaattg gactggagtc actcaaaatg gaacaagctc tgcttgcaaa
aggagatcta ataacagttt ctttagtaga ttgaattggt tgacccactt aaaattcaaa
tacccagcat tgaacgtgac tatgccaaac aatgaaaaat ttgacaaatt gtacatttgg
ggggttcacc acccgggtac ggacaatgac caaatcttcc tgtatgctca aacatcagga
agaatcacag tctctaccaa aagaagccaa caaactgtaa tcccgaatat cggatctaga
cccagggtaa ggaatatccc cagcagaata agcatctatt ggacaatagt aaaaccggga
gacatacttt tgattaacag cacagggaat ctaattgctc ctaggggtta cttcaaaata
cgaagtggga aaagctcaat aatgagatca gatgcaccca ttggcaaatg caattctgaa
tgcatcactc caaatggaag cattcccaat gacaaaccat ttcaaaatgt aaacaggatc
acatatgggg cctgtcccag atatgttagg caaaacactc tgaaattggc aacagggatg
cgaaatgtac cagaaaaaca aactaga
```

Figure 1. A sample RNA sequence of the H3N2 virus.

```
gcaagcgcat gccatgatgg catgagctgg ttaacaattg gaatttctgg tccagacaat
ggagctgtgg ctgtactaaa atacaacgga ataataactg gaaccataaa aagttggaaa
aagcaaatat taagaacaca agagtctgaa tgtgtctgta tgaacgggtc atgttttacc
ataatgaccg atggcccgag taataaggcc gcctcgtaca aaattttcaa gatcgaaaag
gggaaggtta ctaaatcaat agagttgaat gcacccaatt ttcattatga ggaatgttcc
tgttacccag acactggcat agtgatgtgt gtatgcaggg acaactggca tggttcaaat
cgaccttggg tgtcttttaa tcaaaacttg gattatcaaa taggatacat ctgcagtgga
gtgttcggtg acaatccgcg tcccgaagat ggagagggca gctgcaatcc agtgactgtt
gatggagcaa acggagtaaa agggttttca tacaaatatg ataatggtgt ttggatagga
aggaccaaaa gtaacagact tagaaagggg
```

Figure 2. A sample RNA sequence of the H1N1 virus.

3.2. Pre-processing

The sequences of RNA viruses in the dataset consist of a sequence of letters. If all these letters are analyzed together, the model will be unable to recognize possible mutation sites, since now there are only two site the head and tail of the sequence. To accurately see where a mutation is going to happen,

DOI: 10.54254/2755-2721/4/2023429

it is necessary to pay attention to the details of each part of the sequence. Thus, the sequences should be divided into some small pieces. In this experiment, the RNA sequence is divided into several words of the same length. Each word consists of three letters like the diagram shown in the Figure 3 [12]. Then, to facilitate the input of the sequence into the model, use different number to represent these words. Since there are only four letters A, T, C, and G occurred in the sequence, it can be calculated that there are only 64 different possible combinations of these four letters to build three letter words. So, the 64 different words in the divided sequence can be assign as 64 different numbers. Then, replace words with numbers and use these numbers as the input for the model later. Finally, make all the sequences the same length, since RNA sequence might have different lengths, which means they might consist of different numbers of letters. To reduce the difficulty of modelling, find the sequence with the maximum length in the dataset first and save the value of the length. Then, compare the save length value with the value of the length of all sequences in the dataset. If the length value of the detected sequence is less than the value saved, use random items to pad the sequence to meet the requirement of the length of the sequence. Generally, the lengths of the RNA sequences have not so much difference. Thus, the random items will not have great influence on the extended sequences. So far, the pre-processing for RNA sequences is done. The summarized basic information of the datasets is in the Table 1. It shows there are 987 samples in the H3N2 dataset. Each sequence in the dataset has a length of 961. In addition, there are 4609 samples in the H1N1 dataset. Each sequence has a length of 535.

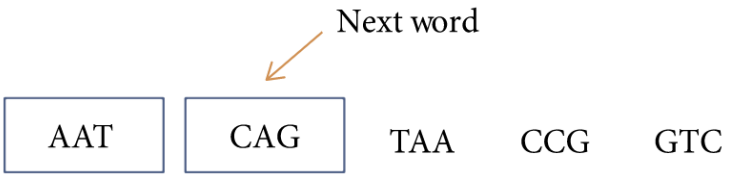

Figure 3. The way to process the RAN sequence.

Table 1. The summarized basic information of the datasets.

	H3N2	H1N1
Number of samples	987	4609
Length of samples	961	535

3.3. Construct a sequence-to-sequence long short-term memory model

Sequence-to-sequence long short-term memory model shows a strong capability to deal with long sequences. This composite model inputs a sequence and outputs a possible predicted sequence. The basic structure of the composite model follows the traditional sequence-to-sequence model, while the difference is its encoders and decoders consist of LSTM layers. So, there are some LSTM layers in the encoder. The LSTM units in the encoder will convert input sequence information into the hidden vector and cell vectors passing to the next LSTM layers in the encoder. During the process of converting input sequence information into hidden vectors and cell vectors, LSTM layers selectively remember and forget information of the sequence to preserve the most important features. The encoder finally passes the hidden states and cell states with highly summarized sequence information to the decoder as the vector form. The formula (1) shows the way to compute the hidden state in the encoder:

$$h(t) = f(W(hh) \times h(t-1) + W(h(x)) \times x(t)) \quad (1)$$

where x(t) refers to the input sequence with length t, h(t) represents the current hidden state, and h(t-1) represents the hidden state before h(t). The decoder's structure is very similar to the encoder, which means the decoder also consists of some LSTM layers. The decoder uses the hidden states and cell states with the vector form from the encoder to produce the predicted sequence. The formula (2) represents the way to compute the hidden state in the decoder:

$$h(t) = f(W(hh) \times h(t-1)) \quad (2)$$

DOI: 10.54254/2755-2721/4/2023429

Here, h(t-1) represents the previous hidden state. Except for the first layer, each LSTM layer gets hidden states and cell states from the previous layer and the final LSTM layer passes the result to a dense layer and uses the Softmax function to get the probability distribution of mutation at each position in the sequence. The formula (3) shows the method to compute the output:

$$y(t) = \text{Softmax}(W(s) \times h(t)) \quad (3)$$

where W(s) is used to represent the weight of its corresponding hidden state. The Softmax function is used to compute the probability distribution. The hidden state and cell state of the first LSTM layer are from the encoder. Then, the predicted result can be inferred. The big picture of the sequence-to-sequence long short-term memory model is demonstrated in Figure 4 [12].

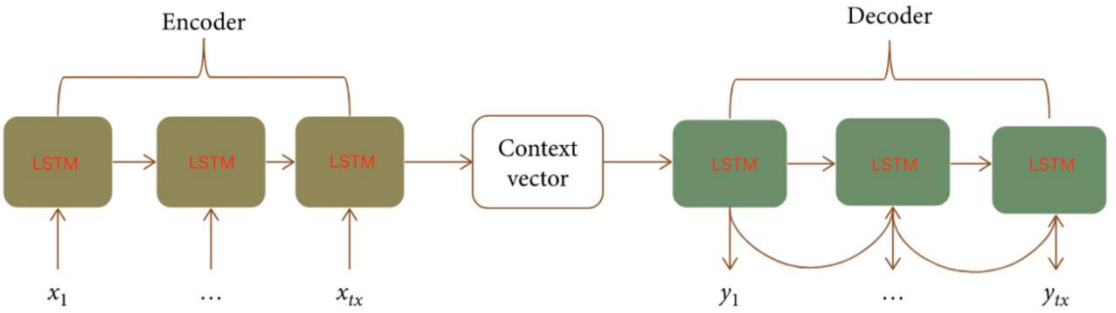

Figure 4. The big picture of the Sequence-to-sequence long short-term memory model.

The final step of constructing a reliable sequence-to-sequence long short-term memory model is configuring the hyper-parameters of the model. The configured best hyper-parameter of the model is shown in Table 2.

Table 2. The hyper-parameter configured in the experiment.

	H3N2	H1N1
Epochs	20	30
Batch	10	10
Activation function	Softmax	Softmax
Training data size	0.8	0.8
Testing data size	0.2	0.2

3.4. Validation

Through the previous training process, the basic structure of the model is determined. The preprocessed validation dataset is inputted into the encoder to obtain the context vector. Then, the decoder will compute results according to the context vector from the encoder. To output the result in a required length, there is a dense layer to deal with the output from the decoder in the final step. Since the input data to the encoder is a sequence of numbers, the output generated by the decoder will be a sequence of numbers as well. To get a readable RNA sequence the integer sequence should be converted back into the corresponding word sequence, which means converting each integer into its corresponding word. Then, spliced all the words together to get the next generation of RNA sequences. The accuracy of the model is the ratio of the number of the correctly next-generation RNA sequence predictions to the total prediction.

4. Results

The model is implemented by the Keras and Tensorflow library in Python 3.8. To evaluate the performance of the model, there are two measures the loss function and accuracy, as shown in Figures 5-8.

For the H1N1 virus dataset, the model has a result of 0.2 loss and 98.3% accuracy. For the H3N2 virus dataset, the model has a result of 0.3 loss and 96.1% accuracy. Both losses of the two models can

DOI: 10.54254/2755-2721/4/2023429

be considered as a low level and the accuracy is very high. So, they should be nice and reliable models. Thus, it can be concluded that these two models are both good enough to provide a reference for predicting the mutation of the viruses selected in the experiment.

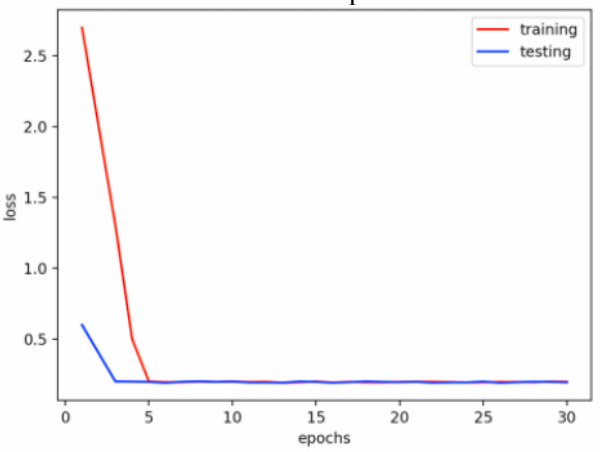

Figure 5. The loss function of the model with H1N1 dataset.

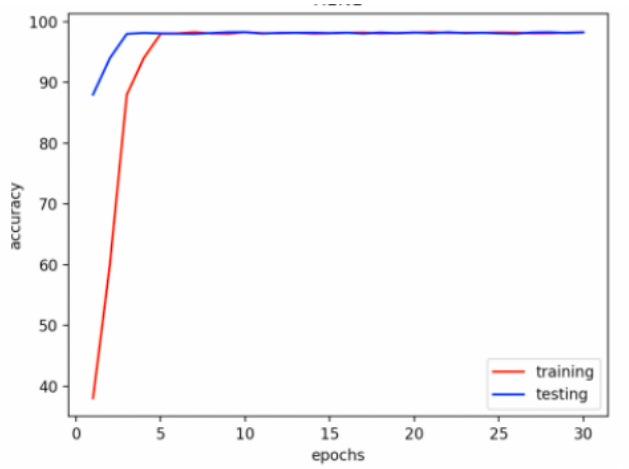

Figure 6. The accuracy of the model with the H1N1 dataset.

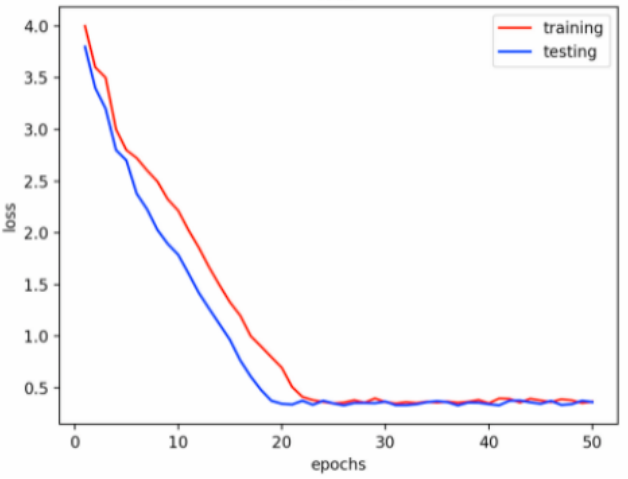

Figure 7. The loss function of model with the H3N2 dataset.

DOI: 10.54254/2755-2721/4/2023429

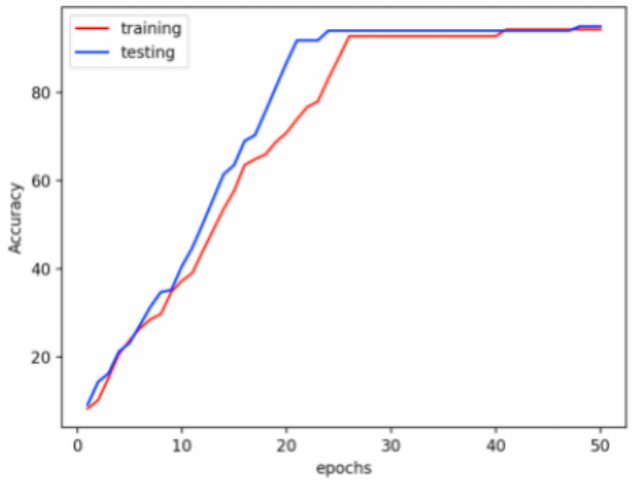

Figure 8. The accuracy of the model with the H3N2 dataset.

5. Discussion

Rapidly and accurately predicting RNA virus mutations can promote the development of vaccines for corresponding RNA viruses. The sequence-to-sequence long short-term memory model in this experiment makes fast and accurate predictions of future mutants based on current RNA sequences, which demonstrated the power of deep learning methods when solving complex problems. One of the hard tasks in this experiment is to process plenty of sequences. For experiments with a lot of sequences requiring to processed, improving the quality of data mining is the key to improving the experimental accuracy. In the experiment, the information of each word is saved in the vector, which greatly improves the speed of information processing and the ability to extract key features of the model. So, good results are not surprising. Nonetheless, this experiment also shows something that needs to be noted in future further research. Firstly, high model accuracy depends on the hyper-parameters of the model. Once the hyper-parameters of the model are set incorrectly, the prediction results may appear a lot of errors. And models for different datasets need to configure different hyper-parameters. Secondly, the two viruses studied in this experiment are both influenza viruses. Whether the method in this experiment is also applicable to other types of viruses such as HIV needs to be further verified. However, anyway, this experiment holds out the promise of predicting virus mutations with little experimental time and financial cost. Although this experiment has some limitations, more datasets and experiments to verify the applicability of the method should be able to dispel people's concerns.

6. Conclusion

Using artificial intelligence methods to predict virus mutation should be a good idea. Both machine learning methods and deep learning methods show great potential in predicting. However, in most past experiments, effectively predicting virus mutation rely on studying several components of the virus. In this experiment, the sequence-to-sequence LSTM model performs well in predicting mutations of the H3N2 virus and H1N1 virus and by recognizing and analyzing the current sequences of RNA alone, the possible future sequence can be predicted by the trained model. This not only indicates that deep learning methods have great potential in predicting the mutation of some viruses but also shows it is possible to predict virus mutation not relying on other components of the virus, like amino acids and proteins. Predicting virus mutation according to the sequences of viruses alone allows people to not study lots of components in the virus, which saves the time and funds of scientific research and improves efficiency. In the future, if there are more virus datasets with more experiments, it is possible to accurately predict the mutations of more kinds of the virus by using deep learning methods and widely using artificial intelligence to predict virus mutations in clinical trials. In addition, finding out the rationales for

DOI: 10.54254/2755-2721/4/2023429

mutation of the virus by using deep learning methods will be greatly helpful for fighting the virus as well, since it might allow people to defeat the virus at its root. To be specific, exploring how the RNA fragments of the viruses choose sites to splice and recombine when they are proliferating by using deep learning methods is looking forward to.

References

[1] Domingo, and Holland, J. J. 1997. RNA virus mutations and fitness for survival. Annual Reviewof Microbiology, 51(1), 151–178. https://doi.org/10.1146/annurev.micro.51.1.151

[2] Cummins, and Badley, A. D. 2015. Can HIV Be Cured and Should We Try? Mayo Clinic Proceedings, 90(6), 705–709. https://doi.org/10.1016/j.mayocp.2015.03.008

[3] Lewin, S. R. 2011. Finding a cure for HIV: will it ever be achievable? Journal of the International AIDS Society. https://link.springer.com/article/10.1186/1758-2652-14-4?error=cookies_not_supportedandcode=921f3430-b41d-45eb-a8c6-7c4d6ec98fb9

[4] Zhang, and Lewin, S. R. 2018. HIV Vaccines and Cure [electronic resource] : The Path Towards Finding an Effective Cure and Vaccine (Zhang and S. R. Lewin, Eds.). Springer Singapore. https://doi.org/10.1007/978-981-13-0484-2

[5] Burrell, C. J., Howard, C. R., and Murphy, F. A. 2017. Pathogenesis of Virus Infections. Fenner and White's Medical Virology, 77–104. https://doi.org/10.1016/b978-0-12-375156-0.00007-2

[6] Cann, A. J. 2012. Genomes. Principles of Molecular Virology, 55–101. https://doi.org/10.1016/b978-0-12-384939-7.10003-1

[7] Tremaglio, C. Z., Barr, J. N., and Fearns, R. 2021. Genetic instability of RNA viruses. Genome Stability, 23–38. https://doi.org/10.1016/b978-0-323-85679-9.00002-7

[8] Payne, S. 2017. Virus Evolution and Genetics. Viruses, 81–86. https://doi.org/10.1016/b978-0-12-803109-4.00008-8

[9] Dugar, P. 2021. Attention — Seq2Seq Models - Towards Data Science. Medium. https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263

[10] I. Sutskever and V. Quoc, 2014, 1409.3215

[11] Saxena, S. 2021. Introduction to Long Short Term Memory (LSTM). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/

[12] Mohamed, T. 2021. "Long Short-Term Memory Neural Networks for RNA Viruses Mutations Prediction. Hindawi.". https://www.hindawi.com/journals/mpe/2021/9980347/

[13] Venugopalan, Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. 2015. Sequence to Sequence -- Video to Text. 2015 IEEE International Conference on Computer Vision (ICCV), 4534–4542. https://doi.org/10.1109/ICCV.2015.515.

[14] Yao, Li, X., Liao, B., Huang, L., He, P., Wang, F., Yang, J., Sun, H., Zhao, Y., and Yang, J. 2017. Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. Scientific Reports, 7(1), 1545–10. https://doi.org/10.1038/s41598-017-01699-z

[15] G. Wu and S. Yan, Amino Acids, vol. 35, no. 2, pp. 365–373, 2008

[16] Xia, Li, W., Li, Y., Ji, X.-L., Fu, Y.-X., and Liu, S.-Q. 2021. A Deep Learning Approach for Predicting Antigenic Variation of Influenza A H3N2. Computational and Mathematical Methods in Medicine, 2021, 9997669–10. https://doi.org/10.1155/2021/9997669.

[17] G. Wu and S. Yan. 2006. Comparative Clinical Pathology, vol. 15, p. 255.

[18] C. L. P. Eng, J. C. Tong, and T. W. Tan. 2014. BMC Medical Genomics, vol. 7, no. 3, pp. 1–11.