Exploring Enterovirus Recombination using Machine Learning

Guang-Wu Chen^{a,b,c,*}, Yu-Nong Gong^{b,c}, Wei-Chung Chen^a

^a Department of Computer Science and Information Engineering, Chang Gung University, 259 Wenhua 1st Road, Guishan District, Taoyuan City 333, Taiwan

^b Research Center for Emerging Viral Infections, Chang Gung University, 259 Wenhua 1st Road, Guishan District, Taoyuan City 333, Taiwan

^c Department of Laboratory Medicine, Linkou Chang Gung Memorial Hospital, 5 Fuxin Road, Guishan District, Taoyuan City 333, Taiwan

*Corresponding Author: gwchen@mail.cgu.edu.tw

Abstract

Enterovirus (EV) genetic recombination has been studied by comparing multiple sequence alignment consisting of both parental and daughter strains. Recently machine learning techniques have also been deployed in sequence analysis. Long short-term memory (LSTM) is a recurrent unit that reads a sequence one step at a time and can model long range relations, and has been used in problems of natural language processing. In this study, we implement LSTM in modeling EV sequences to discover how a daughter strain is composed from recombining its parental strains. We conclude that LSTM can qualitatively capture the recombination pattern of EV, provided that suitable training sequences can be assembled.

Keywords: enterovirus, recombination, machine learning

1. Introduction

Enterovirus (EV), a member of the Picornaviridae family, is a positive-sense, single-stranded RNA virus with approximately 7,400-nt genome. There are over 100 EV serotypes known to infect human, causing various clinical outcomes from mild respiratory illness, hand-foot-mouth disease, to acute hemorrhagic conjunctivitis and aseptic meningitis myocarditis. Some infections could be very serious, leading to neurological complications (e.g., poliomyelitis), pulmonary edema, neonatal sepsis-like disease, acute flaccid paralysis, acute flaccid myelitis, and even fatal. Enterovirus type 71 (also known as EVA-71) caused several outbreaks in Asian countries in the past two

decades. It is one of the widely studied EVs, leading to the most abundant sequences deposited to GenBank (<https://www.ncbi.nlm.nih.gov/nucleotide/>). EVA-71 is also the only serotype adopting homotypic genotype and subgenotype designation to further classify its highly heterogeneous descendants. For example, there are genotype A, B and C, and subgenotype B0 to B5 and C1 to C5. While EV is an RNA virus known to mutate frequently, it frequently utilizes a much more extreme molecular mechanism called recombination to create new viruses by inheriting heterotypic genomic segments from parental strains infecting the same host cell. EV genetic recombination has been studied by comparing multiple sequence alignment consisting of both parental and daughter strains. A number of tools have been developed in analyzing recombination, in particular that SimPlot⁽¹⁾ has been very popular in the research community. Recently machine learning (ML) techniques have also been deployed in sequence analysis. Long short-term memory (LSTM) is a recurrent unit that reads a sequence one step at a time and can model long range relations, and has been used in problems of natural language processing. Liu et al.⁽²⁾ used LSTM to detect DNA modifications on Oxford Nanopore sequencing data. In this study, we implement LSTM in modeling EV sequences to discover how a daughter strain can be composed from recombining its parental strains of different serotype or genotype.

2. Materials and Methods

We chose the full coding sequence from each of the five EV prototypes – EVA-71A (genotype A), EVA-71B

(genotype B), EVA-71C (genotype C), CA4 and CA16, listed in Table 1. Note that the first 3 are of the same serotype with different genotype, while the remaining two are of different serotype. These viruses have been reported to involve in the emerging of a series of EVA-71 C4 subgenotype strains.⁽³⁾ Each of the five prototype sequences was further used as a backbone to simulate for additional 60 genetic variants, where 20 of these carried 1%, 5% and 10% mutations, respectively.

As a result, a total of 305 training sequences consisting of aforementioned 5 EV prototypes, including their simulated mutants of various degrees, were assembled and aligned into a 6,612-nt multiple sequence alignment by MUSCLE⁽⁴⁾. Gaps produced in the alignment were replaced by a 5th character 'N', which accompanies the other 4 bases 'A', 'T', 'C', 'G' to give 5 dimensions (the input dimension) in the training data. The multiple sequence alignment was further divided into smaller and overlapped segments, each of size 500 or 1000.

LSTM models were built using Kera and TensorFlow from training sequences within each of the prescribed window, followed by a sliding-window approach using step size = 100 or 250. The LSTM output dimension was set to 20, and the recurrent neural network (RNN) dimension was set to 32. Two LSTM layers and dropouts were used in constructing

each model, followed by densing using activation function 'softmax' (Figure 1). There were 20% of the training data spared for model validation. We tested our models in the following 3 settings – (1) window size 1000, step size 100, (2) window size 500, step size 100, (3) window size 500, and step size 250. One EVA-71 C4 sequence was used to test these models. The same C4 sequence, together with the 5 prototype sequences mentioned in Table 1, were aligned and analyzed by SimPlot using the same window and step size to benchmark our predicted result.

Figure 1 – A schematic diagram for LSTM model.

Table 1 – Five prototype strains used for generating the needed training data.

Sero-type	Geno-type	Strain name	Year	Accession number
EVA71	A	BrCr	1970	U22521
EVA71	B	MS/7423/87	1987	U22522
EVA71	C	Tainan/4643/98	1998	AF304458
CA4	-	CVA4/SZ/CHN/09	2009	HQ728260
CA16	-	G-10	1951	U05876

3. Results

For training data of window size 1000, 150 epochs were needed for the model to converge within acceptable accuracy and loss, while 300 epochs were required for window size of 500. Setting (2) with bigger window size of 1000 (Figure 2, plotting from x = 500, 600, 700... to 6,100, using a step size of 100) seemed to have a poor prediction comparing to setting (1) of window size 500 (Figure 3). For example, SimPlot predicted 4 recombined segments each from EVA71-C followed by CA4, EVA71-B and CA16, respectively in the bottom of Figure 2. However, LSTM prediction shown in the top half of Figure 2 using bigger window size of 1000 seemed to have wrong classification except for the leading EVA-71C. On the contrary, bottom half of Figure 3 showed a SimPlot result with ambiguous (or mosaic) composition in the middle of the coding sequence. Its corresponding LSTM prediction at the top displayed good qualitative match to the SimPlot result, including the ambiguous region. Using bigger step size 250 in setting (3) comparing to setting (2) of step size 100, on the other hand, deteriorated the prediction (Figure 4). It is clear that all types were mis-predicted except for the EVA-71C in the 5' end. A sample model training accuracy and loss function for one particular 500-nt window were shown in Figure 5. It is noted that the accuracy reached nearly 100% for the training data and at approximately 95% for testing (validation) data with 300 epochs.

Figure 2 – An LSTM result (top) mimicking SimPlot prediction (bottom). The horizontal axis represents 57 1000-nt segments, each overlaps 900-nt from previous segment. Each segment has its own model, and the predicted values were graphed at the center of each 1000-nt segment. For example, the first model produces predicted values of the 5 types at $x = 500$, the second model at $x = 600$, and so on.

Figure 3 – Using window size 500 and step size 100. This is the best setting yielding completely matched result between the two predictions.

Figure 4 – Using window size 500 and step size 250.

Figure 5 – One sample LSTM model training graph for demonstrating excellent accuracy achieved after 300 epochs.

4. Discussion

While SimPlot has been used and helpful in revealing the recombination of enteroviruses in the past, it is extremely important that suitable reference sequences be included to capture the exchange of segment. For example, if CA4 listed in Table 1 did not participate the SimPlot prediction (referring to the bottom of Figure 1), the middle part of the C4 strain will be predicted as EVA-71B only. In other words, the inclusion of prototype sequences used in SimPlot (as well as in our LSTM model) may require a knowledge of how one particular EV recombinant was composed. This is usually not known, unfortunately, demonstrating how such analyses were difficult to handle.

We started with 5 prototypes and 300 simulated mutants as the training data for our LSTM model. We mentioned that the simulation was made from each of the 5 prototype sequences to 1%, 5% and 10% mutation. Before the training, we tested the validity of these data by examining their pairwise identities (data not shown). Chan et al.⁽⁵⁾ analyzed EV genomes and summarized that the average nucleotide divergence between genotypes was 18.99% (17-22%). The inter- and intra-subgenotype average divergences were 12.02% (10-14%) and 3.92% (1-10%), respectively. We found that by including up to 10% mutants into our training data would still allow us to keep adequate separation among serotypes (e.g., CA4 vs CA16, or EVA-71C vs CA16, etc.), and genotypes (e.g., EVA71-B vs EVA71-C). However, we understand that such mutation may not be realistic in actual viral evolution. For example, the mutated nucleotide may introduce stop codon to the EV coding sequences. In the future, we will try using some simulators specifically for

enumerating viral sequences. One example is SANTA-SIM⁽⁶⁾, a software package to simulate the evolution of a population of gene sequences forwards through time.

In this work, we constructed an LSTM model for predicting EV recombination pattern from 5 prototypes. Over 100 EV serotypes/genotypes, how to choose ideal base types remains a challenge. Also we mentioned that our prediction is qualitatively matching to what SimPlot predicted. More efforts will be paid in promoting the quantitative accuracy, in particular that the cross-over site should be pinpointed where recombination occurs.

Although SimPlot prediction was used as a benchmark in assessing the performance of our LSTM model, whether it can completely serve as a golden standard remains to be discussed. As one can see from Figures 2 to 4, different settings for window size and step size can alter the recombination pattern for SimPlot. It will be desired to find other benchmarks that can be used to validate our models.

5. Conclusions

We conclude that LSTM can qualitatively capture the recombination pattern of EV at a setting of window size 500 and step size 100, provided that suitable training sequences can be obtained. Further fine-tuning the training data selection and model parameters may improve the quantitative prediction, which is important for locating the recombination cross-over site.

Acknowledgment

The authors thank to the grant support from the Ministry of Science and Technology, Taiwan (grant no. 107-2221-E-182-064-MY2). We also thank to Mr. Sheng-Da Lee for his discussion in developing the LSTM model.

References

- (1) Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol.* 1999 Jan;73(1):152-60.
- (2) Qian Liu, Li Fang, Guoliang Yu, Depeng Wang, Chuan-Le Xiao & Kai Wang. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications* volume10, Article number: 2449 (2019)
- (3) Zhang Y, Tan X, Cui A, Mao N, Xu S, Zhu Z, Zhou J, Shi J, Zhao Y, Wang X, Huang X, Zhu S, Zhang Y, Tang W, Ling H, Xu W. Complete genome analysis of the C4 subgenotype strains of enterovirus 71: predominant recombination C4 viruses persistently circulating in China for 14 years. *PLoS One.* 2013;8(2):e56341. doi: 10.1371/journal.pone.0056341.
- (4) Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, (5) 113
- (5) Chan YF, Sam IC, AbuBakar S. Phylogenetic designation of enterovirus 71 genotypes and subgenotypes using complete genome sequences. *Infect Genet Evol.* 2010 Apr;10(3):404-12.
- (6) Abbas Jariani, Christopher Warth, Koen Deforche, Pieter Libin, Alexei J Drummond, Andrew Rambaut, Frederick A Matsen IV, Kristof Theys. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evolution*, Volume 5, Issue 1, January 2019.