EC Neurology Special Issue - 2021

Analysis SARS-CoV-2 Genomes of G20 Areas on Phylogeny Tree, t-SNE based on Machine Learning

Renyang Liu, Mu Qiao, Alima, Jeffrey Zheng* and Wei Zhou

Yunnan University, Kunming, China

***Corresponding Author:** Jeffrey Zheng, Yunnan University, Kunming, China.

Received: August 18, 2020; **Published:** January 27, 2021

Abstract

The new coronavirus disease (COVID-19) was outbroken earlier in Wuhan, and the disease spread rapidly from multiple resources of different countries through-out the world. COVID-19 has caused millions of diagnosed people worldwide, causing a large number of deaths, and posing a severe threat to public health in countries around the world. Facing this urgent situation, In-depth research on the emerging SARS-CoV-2 to understand the related pathogenic mechanism and epidemiological characteristics is urgent. This type of activity would be useful to determine its origin to formulate effective prevention and treatment strategies for affected patients. This paper adopts t-SNE based on machine learning to draw a phylogenetic tree from collected genomic sequences to analyze samples of G20 countries. The phylogenetic tree of generating mechanism was described, and intermediate results were illustrated. The results of this research shown that viruses in many countries have similar or similar relationships among the gene sequence.

Keywords: *COVID-19; SARS-CoV-2; Feature Extraction; Machine Learning; Gene Sequence; t-SNE; Phylogenetic Tree*

Introduction

Coronavirus belongs to the Nestoviruses, Coronaviridae and Coronavirus genus. It is a type of RNA virus with an envelope and a linear single-stranded genome. It is a large class of viruses that widely exist in nature. Certain coronaviruses can infect humans and cause diseases, such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS), whose symptoms can range from a common cold to severe lung infections.

The coronavirus outbroken earlier in Wuhan in December 2019 is a virus strain that has not been previously discovered in humans and was named by the WHO as the 2019 novel coronavirus "2019-nCoV" (2019 novel Coronavirus). On February 11, 2020, the virus was named "SARS-CoV-2" (Severe Acute Respiratory Syndrome Coronavirus 2) by the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses. At the same time, the disease caused by the virus infection was named "COVID-19" (Corona Virus Disease 2019) by WHO. 2019-nCoV is a new member of the severe acute respiratory syndrome coronavirus family (SARS-CoV) and is labeled SARS-CoV-2 [7]. In the following days, from Europe to North America, its fatal impact is threatening the entire world. According to the latest update data of the World Health Organization (WHO) so far, more than 2 million people have been diagnosed in more than 200 countries and regions across the country. According to WHO[1], in these cases, about 200,000 confirmed cases died. The new

[1]https://covid19.who.int/

coronavirus (2019-nCoV) meets the definition of all epidemics actively. COVID-19 manifests as fever, sore throat and pneumonia and accompanying severe acute respiratory distress symptoms [6].

Although researchers from worldwide have invested lots of time and energy in this work, the relationship of COVID-19 in different regions has not formed a conclusion yet.

Researchers, however, have proposed many classification algorithms based on machine learning. Most of these algorithms are unsupervised algorithms, which are very suitable for the classification of viruses in various countries, to obtain a rough classification of viruses in various countries. Therefore, in this study, we tried the t-SNE algorithms in machine learning to cluster the gene sequences of G20 countries and construct a phylogenetic tree to show the virus' evolution. From the experimental results, we can see that the virus sequences in these countries can be divided into various major categories, that is to say, the sources of viruses in countries classified by machine learning in the same category may be the same.

Here, we describe our work as follows:

- Exploring the different k of k-mers as virus characteristics,
- We use a variety of machine learning-based algorithms t-SNE to cluster the G20 countries' new coronavirus gene sequences to explore the relationship between viruses in each country,
- Attempt to show the connection and mutation of viruses among countries through phylogenetic tree.

The rest of the paper is composed as follows: Part 2 introduces the materials and methods of our work. Next, we show the experiments and discussion in Part 3. Finally, Part 4 briefly introduces our research conclusions.

Materials and Methods

In this section, we will give a detailed introduction to the materials and methods used in our work.

Datasets

All datasets use in this work from various open-source genomic banks CNCB[2] and GISAID[3] and we cleaned the data. The virus' information are showing in table 1.

No.	Name	Id	Collected Date	Locality
1	Australia.fasta	EPI_ISL_420006	2020-03-24	Australia
2	Belgium.fasta	EPI_ISL_420432	2020-03-23	Belgium
3	Brazil.fasta	EPI_ISL_427306	2020-03-10	Brazil
4	Canada.fasta	EPI_ISL_413014	2020-01-25	Canada
5	Chile.fasta	EPI_ISL_414580	2020-03-05	Chile
6	China.fasta	EPI_ISL_414692	2020-02-25	China
7	England.fasta	EPI_ISL_414500	2020-03-04	England
8	France.fasta	EPI_ISL_414500	2020-03-23	France
9	Germanv.fasta	EPI_ISL_2414521	2020-03-02	Germany
10	India.fasta	EPI_ISL_424365	2020-03-17	India
11	Italy.fasta	EPI_ISL_419254	2020-03-23	Italy
12	Japan.fasta	EPI_ISL_419307	2020-03-20	Japan
13	Mexico.fasta	EPI_ISL_424673	2020-03-12	Mexico
14	Russia.fasta	EPI_ISL_420080	2020-03-18	Russia
15	SaudiArabia.fasta	EPI_ISL_416522	2020-03-10	Saudi Arabia
16	Singapore.fasta	EPI_ISL_420111	2020-03-12	Singapore
17	SouthAfrica.fasta	EPI_ ISL421575	2020-04-01	South Africa
18	SouthKorea.fasta	EPI_ISL_413516	2020-02-27	South Korea
19	Tukey.fasta	EPI_ISL_424366	2020-03-17	Turkey
20	USA.fasta	EPI_ISL_424353	2020-04-02	USA

Table 1: *The information about virus' gene sequence.*

[2]https://bigd.big.ac.cn/ncov/release_genome

[3]https://www.gisaid.org/epiflu-applications/next-hcov-19-app/

t-SNE

t-SNE (t-distributed stochastic neighbor embedding) [3] is a machine learning algorithm for dimensionality reduction. Laurens van der Maaten proposed it and Geoffrey Hinton in 2008 based on SNE [4] Proposed. t-SNE is a non-linear dimensionality reduction algorithm, which is very suitable for high-dimensional data reduction to 2 or 3 dimensions for visualization. Besides, it has meaning when the data is marked, which can clearly show the clustering status of the input data. The main idea is to use conditional probabilities to represent the distance of high-dimensional distribution points, as well as low-dimensional distribution points. As long as the conditional probabilities of the two are very close (training with relative entropy, so labels are needed), it means that the points of the high-dimensional distribution have been mapped to the low-dimensional distribution.

In this article, we use the t-SNE algorithm to cluster viral gene sequences as the following two steps:

- Extract the feature vectors of virus genes from G20 countries.
- Using the t-SNE algorithm to cluster the extracted feature vectors.

Features extraction

As we all known, the gene sequence is composed of four essential elements {"A", "T", "C", "G"}. The length of each COVID-19 virus gene sequence is about 30,000. In order to clustering G20's virus gene sequence with t-SNE, we must extract the corresponding features of each sequence first.

Before that, we first need to know the concept of a proper noun, mer (monomeric unit, mer), which means monomer unit in the field of molecular biology. Units commonly used in nucleic acid sequences represent nt or bp. For example, 100 mer DNA represents a single-strand length of 100 nt or a double-strand length of 100 bp. The k-mer [5] refers to dividing the nucleic acid sequence into a string containing k bases, that is, iteratively selecting a sequence of length k bases from a contiguous nucleic acid sequence. If the length of the nucleic acid sequence is L, k-mer If the length is k, then L-K + 1 k-mers can be obtained.

In our work, we use the number of k-mers as the characteristics of our gene sequence, but unlike other methods of k-mer, we will count the repeated k-mers only once. We need to explore which k-mers are the most important components of gene sequences. Therefore, we designed an experiment to explore the relationship between k-mers of different lengths and viral gene sequences. Finally, we found that the length k satisfying $5 \le k \le 40$ is more appropriate.

Clustering use t-SNE

After extracting the gene feature, we use the number of mer types minus the average and average of the corresponding mer number respectively as the feature vector of each virus gene. Then use the t-SNE algorithm to cluster and display the viral sequences of G20 countries.

Phylogenetic tree

A phylogenetic tree [1] or evolutionary tree is a kind of tree structure diagram commonly used to express the genealogical relationship of species. At the molecular level, the distance between kinship is usually expressed by differences in DNA (or protein) sequences.

There are many ways to build a phylogenetic tree and we use the distance method in our work. The distance-dependent method means that the evolutionary distance of the two sequences determined the topological shape of the phylogenetic tree. The length of the clade branch represents the evolutionary distance.

$$d(x,y) = \sqrt{(x_1 - y_x)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

Specifically, we use UPGMA (Unweighted pair group method with arithmetic mean) based on distance method in our work, where distance refers to Euclidean distance (Equation 1) [2]. The detailed steps are as follows:

- We first generate a "distance matrix" by comparing two gene sequences, and then calculate the gene distance of each pair of sequences. In short, the number of two sequences that do not match (of course, the actual calculation is more important than this Much trouble).
- Then, using the pairwise alignment distance matrix to estimate the two sequences with the shortest distance. These two sequences form the two clades of the evolutionary tree. And then, the distance matrix between these two alignments restarts to find the two closest sequences. However, different from before, the two most similar sequences are connected to the tree by one node at this time. And so on until the end.
- Draw the tree according to the distance with Biopython[4].

Experiments and Discussion

Find the suitable k of k-mers

The results in figure 1 shows that when the k of k-mers of the subsequence is too short, the number of types of the mer is particularly poor. For example, when k equals 1, there are only four types of mer. The k is too large, however, the types of mer are close to the length of the gene sequence. For example, when k equals 100, the number of mer is close to 3000. That is to say, when k is less than or greater than a certain length, it has almost no effect on the division of different genes for each country, and all the polylines will tend to be stable. Moreover, in order to observe more clearly, we subtracted the average of the number of categories corresponding to each mer. The result shows in figure 2.

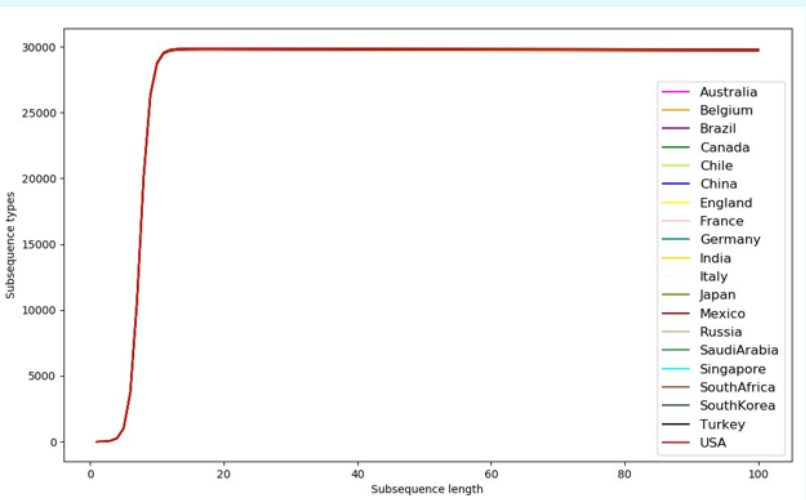

Figure 1: *The relationship between different k and the mer types. The abscissa is the length k of mer, and the ordinate is the various numbers of the mer.*

[4]https://github.com/biopython/biopython.github.io

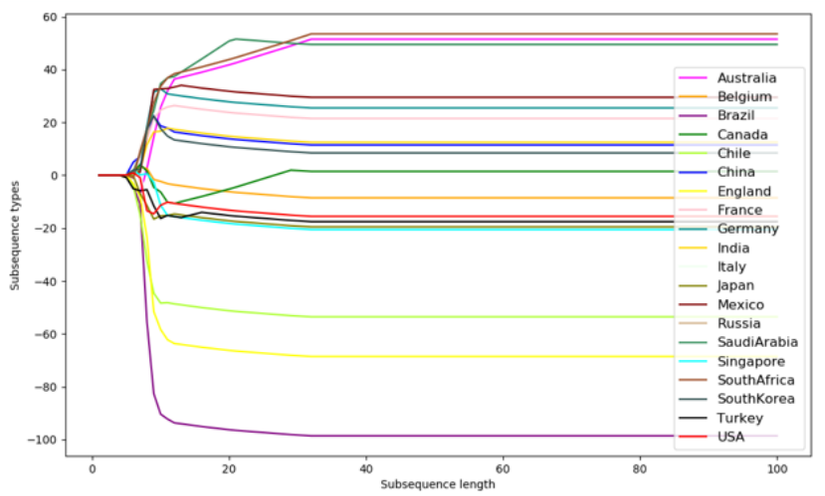

Figure 2: *The relationship between different k and the mer types after subtracting the average value. The abscissa is the length k of mer, and the ordinate is the various numbers of the mer.*

According to the experimental results, we can see that when the length satisfy 5 ~ 40, where k is the length of the mer, the types of mer become more evident, so we using 5 ≤ k ≤ 40 as the length of the mer to extract feature of virus sequence, and carry out the next exploration experiment.

Clustering with t-SNE

The clustering results with t-SNE of the viral gene sequences of G20 countries show in figure 3 and 4. We can see that the t-SNE algorithm clustered the viral gene sequences into at least eight major categories. For example, it clusters France, Germany and Mexico into one category, which means that their virus sources may be the same. Then, the United States, Singapore, Belgium and Turkey can cluster into another. This phenomenon indicates that their virus source may be the same. Japan's virus gene sequence and other countries have not been clustered, but have become a cluster alone, indicating that Japan's virus source maybe It is different from other

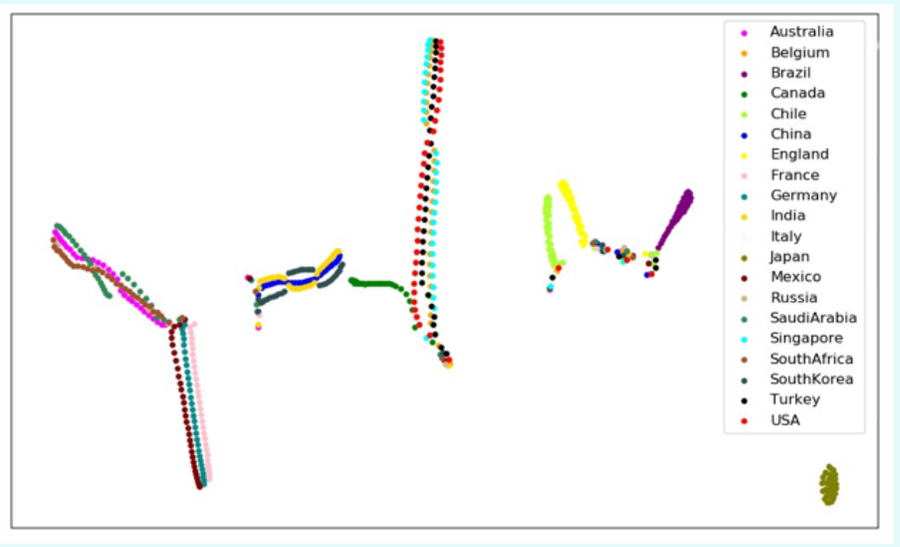

Figure 3: *The clustering results by minus the minimum of corresponding subsequence types.*

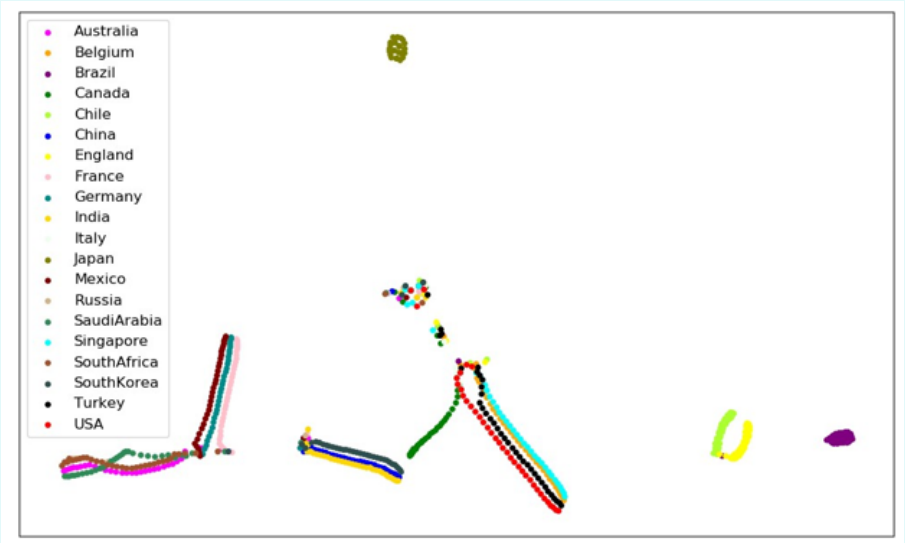

Figure 4: *The clustering results by minus the average of corresponding subsequence types.*

G20 countries.

The G20's phylogenetic tree

Figure 5 is a phylogenetic tree of virus sequences in G20 countries, which shows the evolution of viruses in G20 countries. It is

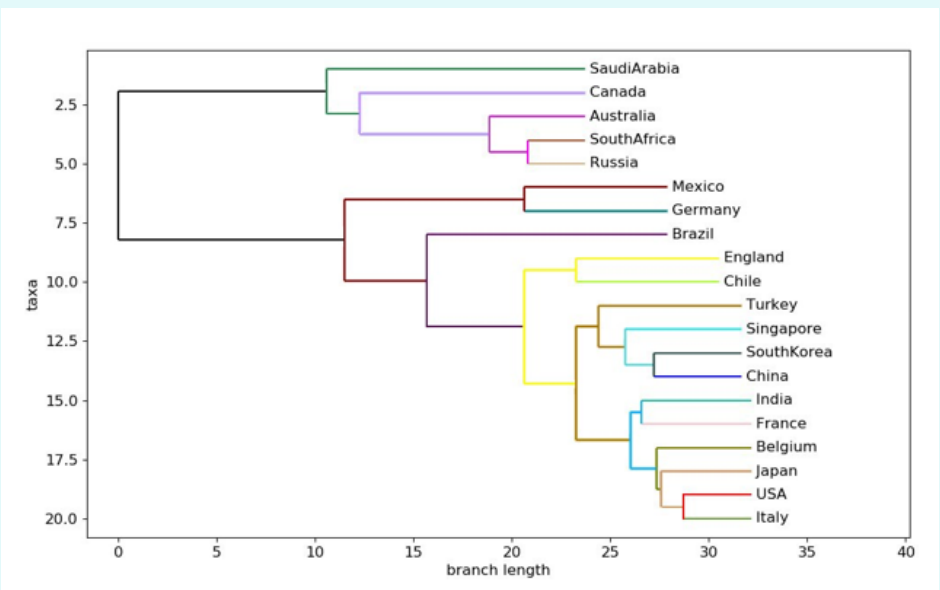

Figure 5: *Phylogenetic tree of viral gene sequences in G20 countries.*

worth mentioning that no matter how we disturb the order of input data, the results we get are the same. This result shows that this phylogenetic tree is relatively stable and can represent the evolution of the gene sequences of G20 countries.

Conclusion

This paper analyzes the gene sequence of the COVID-19 in G20 countries and does many experiments. First, we propose to use a method based on k-mers to extract the features of each virus sequence, select a suitable range of k, and use the mer types as the feature vector of each virus gene. Then, we cluster the virus sequences of various countries by using the machine learning algorithm t-SNE. Finally, we draw the phylogenetic tree of viruses in the G20 countries with subsequences as virus characteristics. Our work found that the viruses of the G20 countries can divide into eight categories roughly. Furthermore, the phylogenetic tree shows that the genetic viruses of each country have a common source and also have their characteristics.

Funding

This work was supported by the NSFC (61762089, 61663047, 62041213), the Science and Technology Innovation Team Project of Yunnan Province (2017HC012), and the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002).

Bibliography

1. John P Archer and David L Robertson. "CTree: comparison of clusters between phylogenetic trees made easy". *Bioinformatics* 23.21 (2007): 2952-2953.
2. Bruce L Golden and Michael O Ball. "Shortest paths with euclidean distances: An explanatory model". *Networks* 8.4 (1978): 297-314.
3. GE Hinton. "Visualizing high-dimensional data using t-sne". *Journal of Machine Learning Research* 9.2 (2008): 2579-2605.
4. Geoffrey E Hinton and Sam T Roweis. "Stochastic neighbor embedding". In Suzanna Becker, Sebastian Thrun and Klaus Obermayer, editors, Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada] MIT Press (2002): 833-840.
5. Pavel A Pevzner., *et al*. "An eulerian path approach to dna fragment assembly". *Proceedings of the National Academy of Sciences of the United States of America* 98.17 (2001): 9748-9753.
6. Samar Salman and Mohamed Salem. "Routine childhood immunization may protect against covid-19". *Medical Hypotheses* 140 (2020): 109689.
7. Sheng Zhang., *et al*. "The novel coronavirus (sars-cov-2) infections in china: prevention, control and challenges". *Intensive Care Medicine* 46 (2020): 591-593.