Autoencoder-based representation learning on SARS-CoV-2 RNA genome sequences

Author: Dong Liang

E-mail: ldifer@gmail.com

Background

The novel coronavirus disease (COVID-19) started in late 2019 has developed into a global pandemic, posing an immediate and ongoing threat to the health and economic activities of billions of people today. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes COVID-19, is characterized by rapid and efficient individual to individual transmission with a range of clinical courses including severe acute respiratory distress syndrome, viral pneumonia, mild upper respiratory infection (URIs) and asymptomatic carriers [1]. Covariates associated with worse outcome include hypertension, diabetes, coronary heart disease and older age. [1] Study on COVID-19 cases on the Diamond Princess cruise ship in Japan estimates the proportion of asymptomatic patients to be 17.9% (95% CrI: 15.5-20.2%)[2]. All these present great challenges for prevention and control of the COVID-19 transmission.

There are clear evidences that the SARS-Cov-2 is evolving rapidly. A recent phylogenetic network analysis of 160 SARS-Cov-2 genomes identified three central variants based on amino acid changes [3]. Yet, Tang *et al* found that two SNPs in strong linkage disequilibrium at location 8,782 (orf1ab: T8517C, synonymous) and 28,144 (ORF8: C251T, S84L) can form haplotypes that classified SARS-CoV-2 viruses into two major lineages (L and S types) [4]. Mutations also frequently occur in the receptor-binding domain (RBD) in the spike protein that mediates infection of human cells [5]. An recent analysis of the viral genomes of 6,000 infected people identified one mutation (named D614G) in the spike protein to be associated with increased virus transmissibility [6]. Obviously, the dynamic evolution of virus genome would have important effects on the spread, pathogenesis and immune intervention of SARS-CoV-2.

Machine learning methods have been successfully applied to classify different types of cancer and identify potentially valuable disease biomarkers [7-14]. In addition, the convolutional neural networks (CNNs) has been developed into the method of choice for medical images recognition and classification. Its special convolution and pooling architectures and parameter sharing mechanism make it computationally more efficient compared to the traditional fully connected neural networks. Albeit with its great popularity in various computer vision tasks, the CNN is less commonly employed in the field of genome sequence analysis. This study attempted to use the state-of-the-art CNN-based autoencoder and perform representation learning on 3161 full-length RNA genome sequences of SARS-Cov-2 collected from across various U.S. states and the world. The model prototype developed in this study could serve as a first step in developing disease risk scoring system in the future.

Representation learning

Virus sequence dataset
The SARS-Cov-2 genome sequence data used in this analysis were obtained from the NCBI GenBank. The NCBI virus repository currently contains 3161 full-length nucleotide sequences of SARS-CoV-2 collected since December 2019 . Each genome is about 30 kb in size, with detailed information about virus collection dates and locations. All the virus sequence data are available from the NCBI virus sequences for discovery platform through the severe acute respiratory syndrome coronavirus 2 data hub

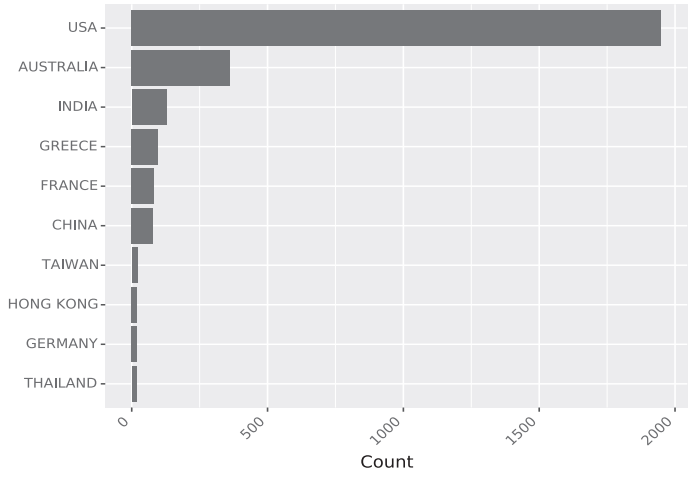

Figure 1: *The top 10 countries/regions with SARS-Cov-2 genome sequences recorded in NCBI virus repository*

(https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049.) As is shown in **Fig. 1:** NCBI virus repository currently records SARS-Cov-2 sequence data from 37 countries/regions around the world, of which the top 10 are USA, Australia, India, Greece, France, China, Taiwan, Hong Kong, Germany and Thailand.

CNN-based autoencoder
The autoencoder was built based on a 2D convolutional neural network with architectures of a range of combinations of convolutional and activations layers. It turns out the best performance was achieved using the simple settings as shown below. The 2D convolutional neural network architecture did yield superior classification performance as compared to the 1D CNN (data not shown).

- Filter size: 5 x 5
- No. of channels: 1, 32, 64, 64, 32, 1
- Batch size: 32
- Activation: ReLU
- Convolutional/deconvolutional layers + Sigmoid layer

The length of the SARS-CoV-2 RNA genome

It can be shown that the length of the SARS-CoV-2 RNA genome has been changing over time. Coronavirus is an enveloped, positive-stranded RNA virus. Among the seven coronavirus subtypes that can infect humans, SARS-CoV-2 belongs to the betacoronavirus lineage B that can severe acute respiratory syndrome in humans [19, 20]. Our preliminary analysis shows that the length of the RNA genome of SARS-CoV-2 has rapidly changed from December 2019 to May 2020 (**Fig. 2**). Although the biological consequences of viral genome changes are yet to be elucidated, these timestamped viral sequences may provide useful clues for evaluating the evolution of SARS-CoV-2 and disease severity.

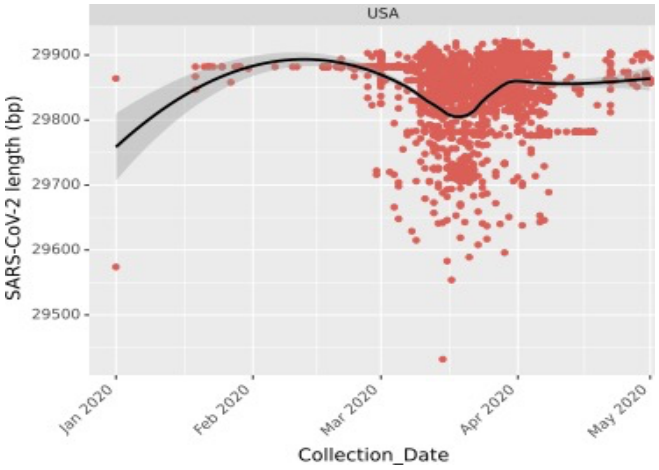

Figure 2: *Length change in RNA genomes of SARS-CoV-2 over time. The scatter plot shows the diversity of the virus genome length from March to April corresponding to the time period when the COVID-19 outbreak unfolded in USA.*

Unsupervised classification of the spreading virus

In order to gain insights into discriminating features of the SARS-Cov-2 currently spreading in different locations, we built an autoencoder-based representation learning model that attempts to classify the types of viruses based on their genome sequences. Additionally, we selected four typical countries - Australia (Oceania), China (Asia), Germany (Europe) and the United States (America), and applied the autoencoder classifier to classify the types of epidemic-causing viruses in these countries between December 2019 and May 2020. Based on the sequence information of the viruses recorded in NCBI virus repository, our unsupervised classification successfully yielded multiple meaningful structures of SARS-CoV-2 clusters (**Fig. 2**).

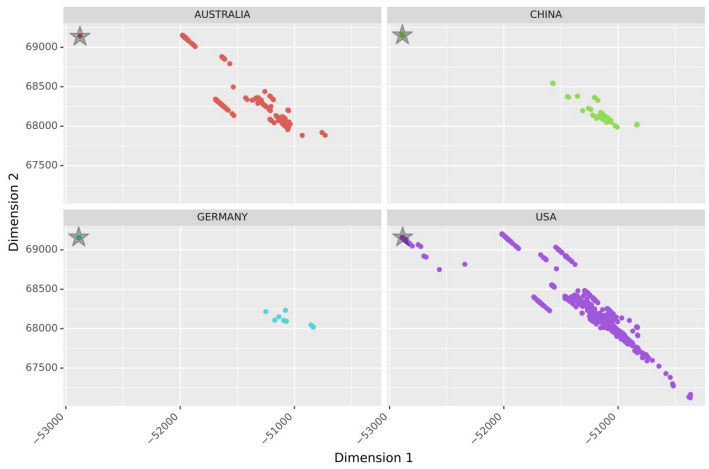

Figure 3*: Representation learning based classification of RNA genomes of the spreading SARS-CoV-2 in four major countries. The stars at the top left corner indicate where locates the reference genome of the virus (Accession ID: NC_045512).*

Our analysis showed that the virus types in China and Germany are relatively homogeneous, whereas those in the United States and Australia presented diversified patterns (**Fig 2**).. In particular, the 'long tail' in its lower right corner is intriguing, as this pattern can only be observed in the United States. These findings are consistent with a previous phylogenetic network

analysis that identified nearly half of the currently documented virus subclusters are mainly in the United States and Australia [3]. In addition, our analysis also demonstrates a clear departure of multiple virus subclusters from its reference genome (indicated as ★ in **Fig. 3**), suggesting that the SARS-Cov-2 is undergoing aggressive evolution.

Discussion

To the best of our knowledge, this is the first study to use SARS-Cov-2 genome sequence data to develop a machine learning model for unsupervised classification of the spreading viruses. In this study, we explored the application of CNN-based autoencoder in processing genome sequence data for classification problems. We showed that convolutional neural network with shift invariant filters are powerful tools for generating discriminating features extracted from the viral sequences.

There are several limitations to this study. First, in this study, the dataset used to perform representation learning included only incomplete virus sequences uploaded between the December 2019 and May 2020 from across the world, although it's a time period coinciding with the onset of outbreaks in many countries. Second, we were not able to track the evolution of virus types over time in these countries, since the vast majority of the viral sequences catalogued by the NCBI virus repository were collected in the March 2020. Finally, because the COVID 19 pandemic remains an ongoing event, its biological consequences and clinical manifestations are yet to be elucidated, which makes it impossible to fully assess the effectiveness and usefulness of the model in predicting and classifying rapidly evolving viruses that are currently emerging and/or undocumented.

Due to time limits, we have not yet explored many of the areas that could expand the findings obtained in this study. We propose the following directions for future studies: 1) Perform unsupervised classification on viral genome sequence data using other traditional machine learning models (e.g. principle component analysis, t-SNE, UMAP, *etc)* and make prediction based on different scenarios; 2) Correlate classification results with clinical phenotypes and/or viral transmissibility, and further construct symptom-specific, supervised representation learning models and 3) Compare the differences and similarities between the classification results of machine/deep learning models and traditional phylogenetic network analysis. All these directions would not be possible without the support of a wealth of clinical and social tracing/spatial-temporal data.

Reference

1. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395(10229):1054-62. Epub 2020/03/15. doi: 10.1016/S0140-6736(20)30566-3. PubMed PMID: 32171076.
2. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Euro Surveill. 2020;25(10). Epub 2020/03/19. doi: 10.2807/1560-7917.ES.2020.25.10.2000180. PubMed PMID: 32183930; PubMed Central PMCID: PMCPMC7078829.
3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020;117(17):9241-3. Epub 2020/04/10. doi: 10.1073/pnas.2004999117. PubMed PMID: 32269081; PubMed Central PMCID: PMCPMC7196762.
4. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020:nwaa036. doi: 10.1093/nsr/nwaa036. PubMed PMID: PMC7107875.
5. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450-2. Epub 2020/04/15. doi: 10.1038/s41591-020-0820-9. PubMed PMID: 32284615; PubMed Central PMCID: PMCPMC7095063.
6. Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020:2020.04.29.069054. doi: 10.1101/2020.04.29.069054.
7. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep. 2015;5:13087. doi: 10.1038/srep13087. PubMed PMID: 26278466; PubMed Central PMCID: PMCPMC4538374.
8. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001;98(26):15149-54. doi: 10.1073/pnas.211566398. PubMed PMID: 11742071; PubMed Central PMCID: PMCPMC64998.
9. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature. 2006;442(7104):772-8. doi: 10.1038/nature04979. PubMed PMID: 16862119; PubMed Central PMCID: PMCPMC2623244.
10. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311-8. doi: 10.1038/ng1966. PubMed PMID: 17277777.
11. Picardi E, Pesole G. Computational methods for ab initio and comparative gene finding. Methods Mol Biol. 2010;609:269-84. doi: 10.1007/978-1-60327-241-4_16. PubMed PMID: 20221925.
12. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol. 1990;212(4):563-78. doi: 10.1016/0022-2836(90)90223-9. PubMed PMID: 2329577.
13. Boissoneault J, Sevel L, Letzen J, Robinson M, Staud R. Biomarkers for Musculoskeletal Pain Conditions: Use of Brain Imaging and Machine Learning. Curr Rheumatol Rep. 2017;19(1):5. doi: 10.1007/s11926-017-0629-9. PubMed PMID: 28144827.
14. Degroeve S, De Baets B, Van de Peer Y, Rouze P. Feature subset selection for splice site prediction. Bioinformatics. 2002;18 Suppl 2:S75-83. PubMed PMID: 12385987.