AutoCoV: Tracking the Early Spread of COVID-19 in Terms of the Spatial and Temporal Dynamics from Embedding Space by K-mer Based Deep Learning

Inyoung Sung[1+], Sangseon Lee[2+], Minwoo Pak[3], Yunyol Shin[3] and Sun Kim[1,3,4*]

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.

[2] BK21 FOUR Intelligence Computing, Seoul National University, Seoul, Republic of Korea.

[3] Department of Computer Science and Engineering Seoul National University, Seoul, Republic of Korea.

[4] Institute of Engineering Research Seoul National University, Seoul, Republic of Korea.

[+] Equal contributor

* Correspondence: sunkim.bioinfo@snu.ac.kr

The widely spreading Coronavirus Disease Virus (COVID-19) has three major properties: pathogenic mutations, spatial and temporal propagation patterns. We know the spread of the virus geographically and temporally in terms of statistics, that is, the number of patients. However, we are yet to understand the spared at the individual patients level. At this point when COVID-19 is spreading all over the world with new genetic variants such as alpha or beta mutation, one important question is to track the early spreading patterns of COVID-19 from a biological perspective until the virus spreads around world. In this work, we proposed a deep learning method, AutoCoV that can track the early spread of COVID-19 in terms of spatial and temporal dynamics of virus spreading patterns until the full spread over the world in July 2020. AutoCoV utilized information theoretic k-mer filtering to preprocess large genome sequences. Then, the deep learning model AutoCoV extended an auto-encoder network with a classifier module and a center loss objective function to track the spatial or temporal dynamics of virus spread patterns. Performances in learning spatial or temporal dynamics were measured with two clustering measurements and one classification measurement. For annotated SARS-CoV-2 sequences from the National Center for Biotechnology Information (NCBI), AutoCoV outperformed seven baseline methods (three dimension reduction methods, two unsupervised methods and two supervised methods) in our experiments for learning either spatial or temporal dynamics. Furthermore, AutoCoV demonstrated the robustness of the embedding space with an independent dataset, Global Initiative for Sharing All Influenza Data (GISAID). In summary, AutoCoV learns geographical and temporal spreading patterns successfully in experiments on NCBI and GISAID datasets and is the first of its kind that leans virus spreading patterns of from the genome sequences, to the best of our knowledge. We expect that is type of embedding methods will be helpful characterizing fast-evolving pandemics.

Keywords: COVID-19, SARS-CoV-2, Deep Learning, Sequence Embedding, Early Spreading Pattern