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[1,3,4,5,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University
[2]BK21 FOUR Intelligence Computing, Seoul National University
[3]Department of Computer Science and Engineering, Seoul National University
[4]Institute of Engineering Research, Seoul National University
[5]Bioinformatices Institute, Seoul National University

Abstract

- The world-wide spreading Coronavirus Disease Virus (COVID-19) has three major properties: pathogenic mutations, spatial and temporal propagation patterns. Although the virus spreads geographically and temporally in terms of statistics, i.e., the number of patients, the understanding of the spread at the individual patient level is still insufficient.
  - ✓ **Goal: Track the early spreading patterns of COVID-19**
- 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

- As illustrated in Figure 1, our model **AutoCoV** consists of four modules:
  1. **Sequence preprocessing**
     : sequences are preprocessed using k-mer for information theoretic filtering and then two level normalization steps are performed, sequence-level and k-mer feature-level
  2. **Auto-Encoder Network**
     : AutoCoV is extended an Auto-Encoder Network and measured reconstructed performance using mean squared error loss
  3. **Classification Network**
     : since SARS-CoV-2 is difficult to construct well-separated embedding space with standard Auto-Encoder Network, adopted an auxiliary Classification Network and trained by minimizing the cross entropy loss
  4. **Center Loss** [1]
     : in order to learn a compact representation of data belonging to the same class, an additional loss function called center loss is used to minimized the distance between data within the same class and learn compact embedding space

  for learning spatial or temporal dynamics of SARS-CoV-2 sequences by constructing two dimensional (2D) embedding space.

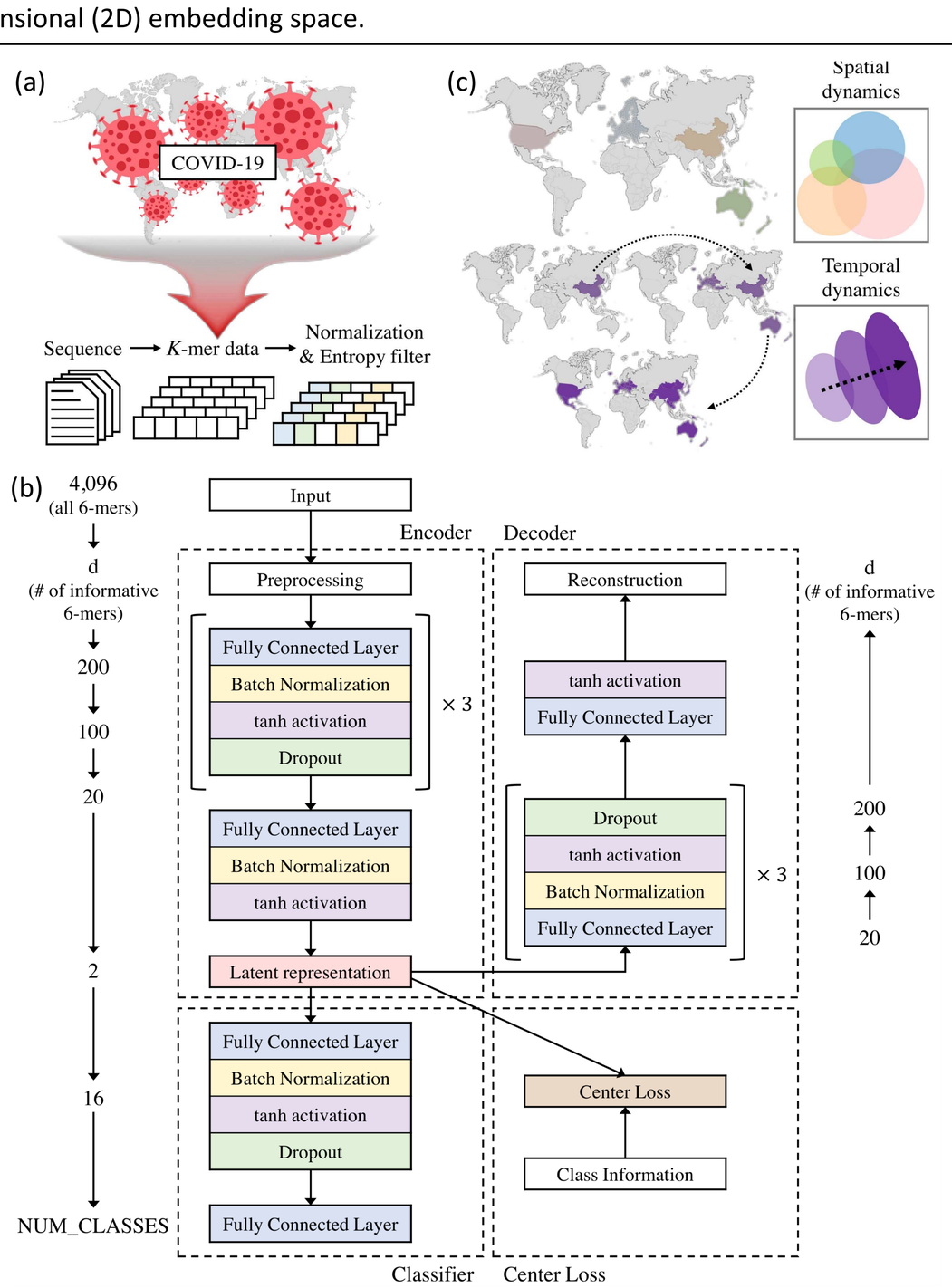

Figure 1. The overall framework of the AutoCoV model. (a) Preprocessing of SARS-CoV-2 sequences. (b) The structure of AutoCoV. (c) The output of AutoCoV.

Motivation

- COVID-19 is wide-spread all over the world with new genetic variants. Once the spread all over the world, it is very difficult to track spreading patterns of COVID-19. Nevertheless, there is no doubt that SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) genome sequences have different characteristics as location-wise and time-wise.
- Beyond the simple statistics, the spread at the level of individual patients can be investigated through an embedding space with spatial and temporal features.
- When external factors such as global lockdown are enforced, it is difficult to investigate spreading potentials of COVID-19 per se, thus we analyzed the early spread pattern of COVID-19 using the SARS-CoV-2 sequences up to July 2020.

Results

- SARS-CoV-2 dataset – NCBI Virus and GISAID
  - Spatial labels: Asia, Oceania, Europe and North America, based on the collected continents
  - Temporal labels: Early, Middle and Late, based on the March 2020
- Performances in learning spatial or temporal dynamics were measures with two clustering measures and one classification measure and were compared with seven baseline methods.
- **Visualization of spreading patterns on spatial or temporal dynamics**: as showed in Figure 2, AutoCoV was the only one that could infer the spreading patterns of SARS-CoV-2 sequence as well as well-separated and clustered each class in both dynamics [2].
  - For spatial dynamics, AutoCoV constructed the space that successfully distinguishes Asia and North America.
  - For temporal dynamics AutoCoV distinguished middle and late time points of SARS-CoV-2.
- **Quantitative measurements**: for sequences from NCBI, AutoCoV outperformed seven baseline methods in our experiments for learning either spatial or temporal dynamics.
  - For spatial dynamics, AutoCoV had at least 1.7-fold higher clustering performance and an F1 score of 88.1%.
  - For temporal dynamics, AutoCoV had at least 1.6-fold higher clustering performances and an F1 score of 76.1%.
- **External dataset validation**: for sequences from GISAID, AutoCoV demonstrated the robustness of the embedding space with an independent dataset.

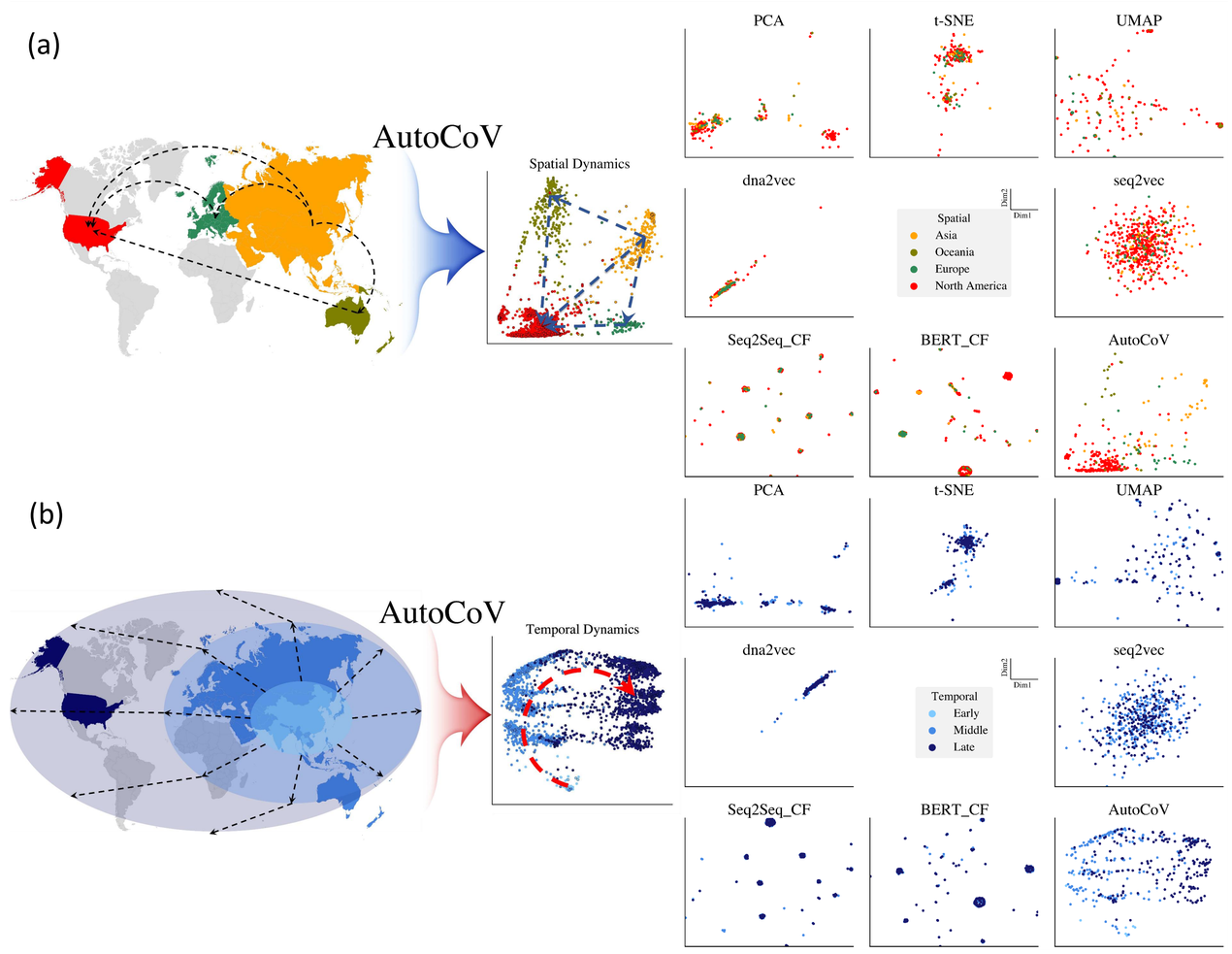

Figure 2. Spreading patterns on each dynamics (a) Spatial and (b) Temporal from Left: AutoCo and Right: Comparison methods.

Conclusion

- **AutoCoV** learns 2D embedding space for modeling the spatial and temporal dynamics of COVID-19 early spreading patterns and is the first of its kind that leans virus spreading patterns form the genome sequences, to the best of our knowledge.
- **Technical contribution**: AutoCoV effectively handle the long-length SARS-CoV-2 genome sequences using information theory and auto-encoder based deep learning models.
- **Biological contribution:** AutoCoV map SARS-CoV-2 sequences to 2D spaces that preserve the spatial or temporal dynamics.
- We expect that comprehensive analysis of statistical methods based on demographic data and this type of embedding method will help characterize the rapidly evolving pandemics

References

[1] Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." European conference on computer vision. Springer, Cham, 2016.
[2] Hadfield, James, et al. "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics 34.23 (2018): 4121-4123.