Embedding Space Alignment of DNA and RNA Space for COVID-19 Severity Prediction Using Deep Learning

Inyoung Sung
BK21 FOUR Intelligence Computing
Seoul National University

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) at the end of 2019 led to a worldwide pandemic characterized by rapid transmission and significant public health effects. As the virus spread, various mutations occurred, and the infected patients showed different symptoms accordingly. According to this, it is shown that the host, which is the affected patient, reacts differently depending on the infected virus, that is, what mutations it has. Therefore, if the virus and the host are analyzed in an integrated way, it will enhance understanding of the disease and make effective severity predictions. In this study, we propose a deep learning-based shared encoder model to integrate viral genome sequences and patient gene expression profiles. This model converted the SARS-CoV-2 DNA sequence into real value data through k-mer analysis, creating real values with mutation information embedded. Using virus-host protein interaction information through biological prior knowledge, shared embedding simultaneously learned the expression values of host proteins corresponding to viral proteins. As a result, the proposed method aligned data from two different regions into the same embedding space, allowing simultaneous prediction of disease severity in both the virus and the patient. In addition, by integrating disparate data, it is expected to provide a comprehensive understanding of how viral mutations affect host cell pathways and immune responses, enhancing knowledge of disease mechanisms and making targeted therapeutic interventions.