Leveraging Singular Value Decomposition(SVD) to Detect SARS-CoV-2 Variants

Wang, Aidan (School: University School of Milwaukee)

As the Coronavirus continues to spread globally, different mutations arise. These variants often take long periods of time until being discovered, resulting in large spikes of infection. This study aims to examine a mathematical approach using singular value decomposition (SVD) to rapidly identify variants. SVD is a fundamental theorem of linear algebra used in Data Science and as a data reduction method in Machine Learning. SVD allows us to write an n x p matrix as a product of three different matrices, A = U Σ VT. The diagonal values in the Σ matrix are known as the singular values. The largest singular value provides us with directions of the maximum variance. This study utilized genomic sequence samples from the GISAID database with a collection date in consecutive months. Due to possible insertions and deletions at some positions, the genomes have different lengths. We have to align the actual genomes with the reference genome(Wuhan patient zero) with Clustal Omega. After the alignment, we compared the genomic sequences with the reference genome and created a binary matrix, with 0 representing the same nucleotide base and 1 representing a difference. The singular values are then calculated using genomic sequences from patients of the four different variants. The experiment result showed the efficacy of the SVD algorithm. Delta and Omicron had a much larger singular value, showing the significant amount of mutations. For the same variant, the smallest singular values started approaching 0 after about 10 sequences. This study demonstrated a viable solution to analyze Covid-19 data and rapidly identify new SARS-CoV-2 variants.