A Novel Alignment-free Genome Sequence Comparison Using Fourier Power Spectrum: Define Distance Metrics Leveraging Singular Value Decomposition

Wang, Aidan (School: University School of Milwaukee)

Sequence alignment tools have traditionally been used to study the evolution of genome sequences by aligning sequences of varying lengths. However, these tools are memory-intensive, time-consuming, and have limited capacity. Alignment-free approaches have been proposed in recent years to work around this limitation by using algorithms such as k-mers word pattern counts and Fourier Transformation. In my previous research, I used Fourier Transformation to retrieve the mean value of the Power Spectrum and added statistical moments to reflect the distribution of Power Spectrum. This algorithm yielded promising results when comparing genome sequence data in the same region. However, the results weren't as consistent in some regions, especially China, in comparison with other regions such as the U.S. and UK. This observation motivated this year's research to (1) explore alternate algorithms to generate an approximation of the Fourier Power Spectrum (2) define distance metrics to measure the similarity/dissimilarity of SARS-CoV-2 sequences across different regions. This new approach using Singular Value Decomposition (SVD) produced more consistent results when analyzing sequence data across various regions. I also discovered that distance metrics using SVD on the spike(S) protein produced similar results in comparison to whole genome sequence which significantly reduced computation time. As sequencing technology continues to advance, an ever-increasing amount of genome sequence data is becoming readily available. The use of alignment-free method offers significant advantages for large-scale genome analysis. The developed method can be adapted for use with any emerging viruses to track mutations and identify new variants that have potential immune evasion.