HTIC
HealthTech Innovation
CONFERENCE

HealthTech Innovation Conference

Zabrze, 10 – 11th October 2022

Julia UZDOWSKA[1], Anna TAMULEWICZ[1]

[1] *Silesian University of Technology, Faculty of Biomedical Engineering, Zabrze, Poland*

ANALYSIS OF THE SARS-COV-2 MOLECULAR SEQUENCES USING BIOINFORMATICS TOOLS

Keywords: SARS-CoV, COVID-19, Phylogenetic trees, Bioinformatics, Amino acid analysis

The SARS-CoV virus (Severe acute respiratory syndrome coronavirus 2) was first detected in December 2019 in the Chinese city of Wuhan, and then spreading rapidly, led to the declaration of a global pandemic, causing more than 500 million cases and more than 6 million deaths. The aim of the study was to analyze the amino acid sequences of the SARS-CoV-2 coronavirus and homologous sequences with the use of bioinformatics tools. In result, phylogenetic trees were created for all selected sequences that illustrated the changes taking place in coronavirus proteins that could lead to the formation of SARS-CoV-2.

Four structural proteins of coronaviruses were selected for analysis: S - forming coronavirus spikes, M - building the membrane, N - occurring in the nucleocapsid and E - being part of the virus envelope. The virus sequences used for analysis were selected from a literature review, and the sequences themselves were found in the publicly available NCBI database. The sequence dataset has also been extended using the NCBI BLAST tool, which enables NCBI searches to find sequences similar to the selected proteins. In result, sequences of 52 viruses were used in the study. The particular focus was on viruses infecting humans, as well as bats, which are perfectly adapted to the role of coronavirus carriers (they spread over long distances, which allows for a wide range of the virus, and usually the infection does not cause dangerous symptoms in them). It is worth noting that bats previously contributed to the coronavirus pandemic - SARS in 2003 and MERS in 2012.

In order to create phylogenetic trees, multiple sequence alignments were created using the progressive method and the PAM210 amino acid substitution matrix. Then, on the basis of the sequences, distance matrices were created and phylogenetic trees were built using two methods: UPGMA and neighbor joining. As a result, eight trees were obtained, two for each protein. The obtained trees were validated using the Bootstrap method for 100 replicates.

The analysis of the obtained proteins led to the discovery of a group among the studied viruses that contained the sequences most similar to the SRAS-CoV-2 sequence, and thus the closest to it. This group includes the bat viruses RaTG13, Rp3, bat-SL-CoVZXC21 and bat-SL-CoVZC45, the pangolin virus, and the SARS-CoV virus. Regardless of the protein tested or the method used, the closest relationship was found for RaTG13, followed by the pangolin coronavirus.