# Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)



Tasneem Ayman Elmetwally

Third Year (Medical Informatics Program)

**Abstract**

This paper reports and analyses genomic mutations in the coding regions of SARS-CoV-2 and their probable protein secondary structure and solvent accessibility changes, which are predicted using deep learning models. Prediction results suggest that mutation D614G in the virus spike protein, which has attracted much attention from researchers, is unlikely to make changes in protein secondary structure and relative solvent accessibility. Based on 6324 viral genome sequences, they create a spreadsheet dataset of point mutations that can facilitate the investigation of SARS-CoV-2 in many perspectives, especially in tracing the evolution and worldwide spread of the virus. The analysis results also show that coding genes E, M, ORF6, ORF7a, ORF7b and ORF10 are most stable, potentially suitable to be targeted for vaccine and drug development

## Introduction

Scientists have been able to obtain genomic sequences of SARS-CoV-2 and have started analysis of these data. Reference genome of SARS-CoV-2 shows that SARS-CoV-2 is an RNA virus having a length of 29,903 nucleotides. Tracing the evolution and spread of the virus is important for developing vaccines and drugs as well as proposing appropriate intervention strategies. Monitoring and analysing the viral genome mutations can be helpful for this task. Due to a strong immunologic pressure in humans, the virus may have mutated over time to circumvent responses of the human immune system. This leads to the creation of virus variants with possible different virulence, infectivity, and transmissibility. This paper reports all point mutations occurring so far in SARS-CoV-2 and presents exemplified implications obtained from the analysis of these mutation pattern data. Four types of mutations, which include synonymous, nonsynonymous, insertion and deletion, are detected. They use 6324 SARS-CoV-2 genome sequences collected in 45 countries and deposited to the NCBI GenBank so far and create a spreadsheet dataset of all mutations occurred across different genes. Eleven protein coding genes of SARS-CoV-2 have been identified, namely ORF1ab, spike (S), ORF3a, envelope (E), membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N) and ORF10. The order of these genes and their corresponding length are illustrated in Fig. 1. In this paper, to evaluate the possible impacts of genomic mutations on the virus functions, they propose the use of the SSpro/ACCpro 5 methods to predict protein secondary structure and relative solvent accessibility. By comparing the prediction results obtained on the reference genome and mutated genomes, we are able to assess whether the detected mutations have the potential to change the protein structure and solvent accessibility, and thus lead to possible changes of the virus characteristics. Because of the functional importance of structural proteins, they only report the prediction results of these proteins in this study.

