# 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)

https://www.nature.com/articles/s41598-021-83105-3#Sec6

**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.



## Related works

Phan16 analysed 86 genomes of SARS-CoV-2 ( <https://www.gisaid.org/> ) and found 93 mutations over the entire viral genome sequences. three of them occurring in the RBD region of the spike surface glycoprotein S, including N354D, D364Y and V367F, with the numbers showing amino acid (AA) positions in the protein.Also it shows three deletions in the genomes of SARS-CoV-2 obtained from Japan, USA and Australia.Another study in17 shows that the SARS-CoV-2 genomes may have undergone recurrent, independent mutations at 198 sites with 80% are of the nonsynonymous type. A SNP genotyping study in18 discovered highly frequent mutations in the genes encoding the S protein, RNA polymerase, RNA primase, and nucleoprotein, and those may be linked to the virus transmissibility and virulence. Tang et al.19 studed 103 genomes of COVID-19 patients and discovered mutations in 149 sites of these genomes. They found that the spike gene S consistently has larger dS values than other genes.Two major strains of the virus have been identified, denoted as L and S, based on two tightly linked SNPs.The L strain is found more prevalent than the S strain among the examined sequences. Korber et al.20 tracked the mutations of spike protein S of SARS-CoV-2 21. They detected 14 mutations in the growing spike protein, especially the mutation D614G that rapidly becomes the dominant form when spread to a new geographical region. Hashimi22 analysed the mutation frequency in the spike protein S of 796 SARS-CoV-2 genomes.64 mutations were found occurring in the S protein sequences taken from multiple countries. It suggests that the virus is spreading in two forms, the D614 form (residue D at position 614 in the S protein) takes 68.5% while the G614 form takes 31.5% proportion of the examined isolates. Koyama et al.23 found several variants of SARS-CoV-2 that may cause drifts and escape from immune recognition by using the prediction results of B-cell and T-cell epitopes in24. Shen et al.12 conducted metatranscriptome sequencing for bronchoalveolar lavage fluid samples taken from 8 patients with COVID-19 and found no evidence for the transmission of intrahost variants as well as a high evolution rate of the virus with the number of intrahost variants ranged from 0 to 51 around a median number of 4. Pachetti et al.27 examined 220 genomic sequences of COVID-19 patients and discovered 8 novel recurrent mutations at nucleotide locations 1397, 2891, 14408, 17746, 17857, 18060, 23403 and 28881. Mutations at locations 2891, 3036, 14408, 23403 and 28881 are mostly found in Europe while those at locations 17746, 17857 and 18060 occur in sequences obtained from patients in North America. A study in28 on 95 SARS-CoV-2 complete genome sequences discovered 116 mutations. Among them, the mutations at position C8782T in the ORF1ab gene, T28144C in the ORF8 gene and C29095T in the N gene are common.

|  |  |
| --- | --- |
| 16 | [Genetic diversity and evolution of SARS-CoV-2 - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1567134820300915?via%3Dihub) |
| 17 | [Emergence of genomic diversity and recurrent mutations in SARS-CoV-2 - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1567134820301829?via%3Dihub) |
| 18 | [Genotyping coronavirus SARS-CoV-2: methods and implications - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0888754320303189?via%3Dihub) |
| 19 | [On the origin and continuing evolution of SARS-CoV-2 | National Science Review | Oxford Academic (oup.com)](https://academic.oup.com/nsr/article/7/6/1012/5775463) |
| 20 | [Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2 | bioRxiv](https://www.biorxiv.org/content/10.1101/2020.04.29.069054v2) |
| 21 | [A short review on antibody therapy for COVID-19 - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2052297520300342?via%3Dihub) |
| 22 | [Emergence of mutations and possible antigenic drift in the surface glycoprotein of SARS-CoV-2 (COVID-19) (authorea.com)](https://www.authorea.com/users/314251/articles/444577-emergence-of-mutations-and-possible-antigenic-drift-in-the-surface-glycoprotein-of-sars-cov-2-covid-19?commit=6066a33a77d398dc70a36f670c43a2517a94e467) |
| 23 | [Pathogens | Free Full-Text | Emergence of Drift Variants That May Affect COVID-19 Vaccine Development and Antibody Treatment (mdpi.com)](https://www.mdpi.com/2076-0817/9/5/324) |
| 24 | [A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2: Cell Host & Microbe](https://www.cell.com/cell-host-microbe/fulltext/S1931-3128(20)30166-9?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1931312820301669%3Fshowall%3Dtrue) |
| 12 | [Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019 | Clinical Infectious Diseases | Oxford Academic (oup.com)](https://academic.oup.com/cid/article/71/15/713/5780800) |
| 27 | [Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant | Journal of Translational Medicine | Full Text (biomedcentral.com)](https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-020-02344-6) |
| 28 | [Genomic characterization of a novel SARS-CoV-2 - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2452014420300960?via%3Dihub) |