# mutation\_scatter\_plot: Tackling codon usage analysis from a different angle

Mokrejš Martin1, Zahradník Jiří1

1 First Faculty of Medicine, Charles University, BIOCEV, Prumyslova st. 595, Vestec 25250, Czechia

Codon usage is commonly used in phylogenetics to unleash evolution of protein-coding regions, in particular, as the ratio between synonymous and non-synonymous changes. We wanted to study multiple sequence alignment of the Covid19 S protein in a vast set of raw NGS reads and determine which codons out of the theoretical 64 are used in every amino acid position of the encoded protein. To our surprise we could not find a tool to achieve that. Polishing the multiple sequence alignment spanning dozens of millions of entries is not possible with conventional tools although mostly just some gaps would need to be introduced here and there. The major obstacle is introducing padding gaps into the reference sequence to facilitate recognition of INSertion events in the sample reads. First, we developed rather simple program *calculate\_codon\_frequencies.py* to count the codons occurring in three columns of the DNA alignment while moving along the reference sequence and keeping ribosome reading-frame of the CDS region and output TSV files with their frequencies. Alternatively, the same tool can provide frequencies of the encoded amino acid residues. Second, we developed *mutation\_scatter\_plot.py* to display the frequencies as scatter plots with interactive bubbles upon mouse hover(). The changes can be color-coded according to e.g. physicochemical properties of the amino acid residues (PAM matrices) or their evolutionary conservation (BLOSUM matrices) or any other color-palette. However, such efforts are a bit naïve as the weights for each amino acid are not within the same minimum-maximum range and thus are not directly comparable. The software is available at <https://github.com/host-patho-evo/mutation_scatter_plot> . This was in part supported by Czech Science Foundation Grant No. 25-17643M: „Unveiling Divergence and Convergence Points in Coronavirus Evolution for Host Receptor Recognition“ and by National Institute of virology and bacteriology (Programme EXCELES, ID Project No. LX22NPO5103).

*Maximum abstract length is ½ of A4 page.*