

Modeling the Correlations of Coronavirus and Codon Using Measures

Biljana Stojanović

*Mathematical Institute SASA, Knez Mihaila 36, 11000 Belgrade, Serbia
e-mail: bstojanovic@mi.sanu.ac.rs*

Saša Malkov

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: sasa.malkov@matf.bg.ac.rs*

Nenad Mitić

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: nenad.mitic@matf.bg.ac.rs*

Miloš Beljanski

*Institute for General and Physical Chemistry, Studentski trg 16, 11000 Belgrade, Serbia
e-mail:*

Gordana Pavlović Lažetić

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: gordana@matf.bg.ac.rs*

Mirjana Maljkovic Ružičić

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: mirjana.maljkovic@matf.bg.ac.rs*

Ivan Čukić

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: ivan.cukic@matf.bg.ac.rs*

Aleksandar Veljkovic

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: aleksandar.veljkovic@matf.bg.ac.rs*

Stefan Kapunac

*Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: stefan.kapunac@matf.bg.ac.rs*

Abstract. The large number of sequenced isolates of the coronavirus family represents a massive sample for various bioinformatics experiments, including research of genomic variability and research of different data modeling techniques. The goal of our research was to use data mining techniques to determine correlation between codon usage and different types of viruses and proteins. The material includes 980,554 isolates with 15,573,303 coding sequences (proteins) of 7 coronavirus types. Material was downloaded from NCBI (26.05.2023.). Individual coding sequences containing ambiguous nucleotide codes were eliminated. As a measure of codon usage various measures (RSUC, ENC, RCBS, codon frequencies, and others) were used. We have tested different classification and clustering algorithms to construct models based on the used measures. The results we obtained showed that codon measures can be used to construct prediction models that predict the type of virus or protein with very high accuracy (from 96.3% to 99.9%). The clustering of the material used led to a separation of records that is very close to the natural clustering by protein type.

Keywords: bioinformatics; data mining; coronavirus; classification; clustering.