Multiple Sequence Alignment of SARS-CoV-2 genes

Antonio Osamu Katagiri Tanaka[1]*

[1] *ITESM, Av. Eugenio Garza Sada 2501 Sur, N.L., Monterrey, Mexico*

Coronaviruses (CoVs) are named after the crown-like spikes present on their surface. The coronaviruses that affect humans were first identified in the 1960s. The human coronaviruses identified so far are: **SARS-CoV** (Severe Acute Respiratory Syndrome Coronavirus); **NL63** (Human Coronavirus NL63 Amsterdam 1); **OC43** (Organ Culture 43); 229E (Human Coronavirus 229E); **MERS-CoV** (Middle East Respiratory Syndrome Coronavirus); and the novel coronavirus, **COVID-19** (Coronavirus Disease 2019) (1). People around the globe typically get infected with human coronaviruses NL63, OC43, and 229E. However, as is in the case of SARS-CoV, MERS-CoV, and COVID-19 some viruses that only affect animals can evolve to infect humans. (2)

This assignment is to discover similar genes within the novel coronavirus. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) protein sequences were downloaded from NCBI GenBank database at https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/. This work includes the following SARS-CoV-2: E, M, N, ORF1AB, ORF3, ORF8, and S. Protein sequences were collected from different places and different times, as detailed in Table S1.

Sequences were grouped by protein and then inputted to three Multiple Sequence Alignment tools (Clustal Omega, MUSCLE, and webPRANK) available in the European Bioinformatics Institute (EBI) website https://www.ebi.ac.uk/services. Subsequently, the alignment tool output was processed with MView (also available in EBI's website) to better visualize the sequences and calculate the percentage coverage (Equation (1)) and percentage identity (Equation (2)) of each sequence. MView calculates percent coverage and percent identity of every sequence with respect to a reference sequence (by default the first row of a sequence alignment). For more information visit the MView API documentation at https://desmid.github.io/mview/index.html .

$$cov = \frac{\text{number of residues in row aligned with reference row}}{\text{length of ungapped reference row}} \times 100 \tag{1}$$

$$pib = \frac{\text{number of identical residues}}{\text{length of ungapped reference row over align region}} \times 100 \tag{2}$$

*E-mail: A01212611@itesm.mx

From the resulted alignments, general similarities between the sequences were discovered. Even though the sampled sequences came from different locations and times, the founded similarities are expected since the aligned sequences correspond to the same gene and the early discovery of SARS-CoV-2 hence the time gap is only of approximately three months.

In order to very the well functionality of the EBI's tools, Middle East Respiratory Syndrome Coronavirus genes (MERS) were added to each sequence set with the expectation of finding differences. Figure 1 and Figure 2 summarize the sequence alignments.

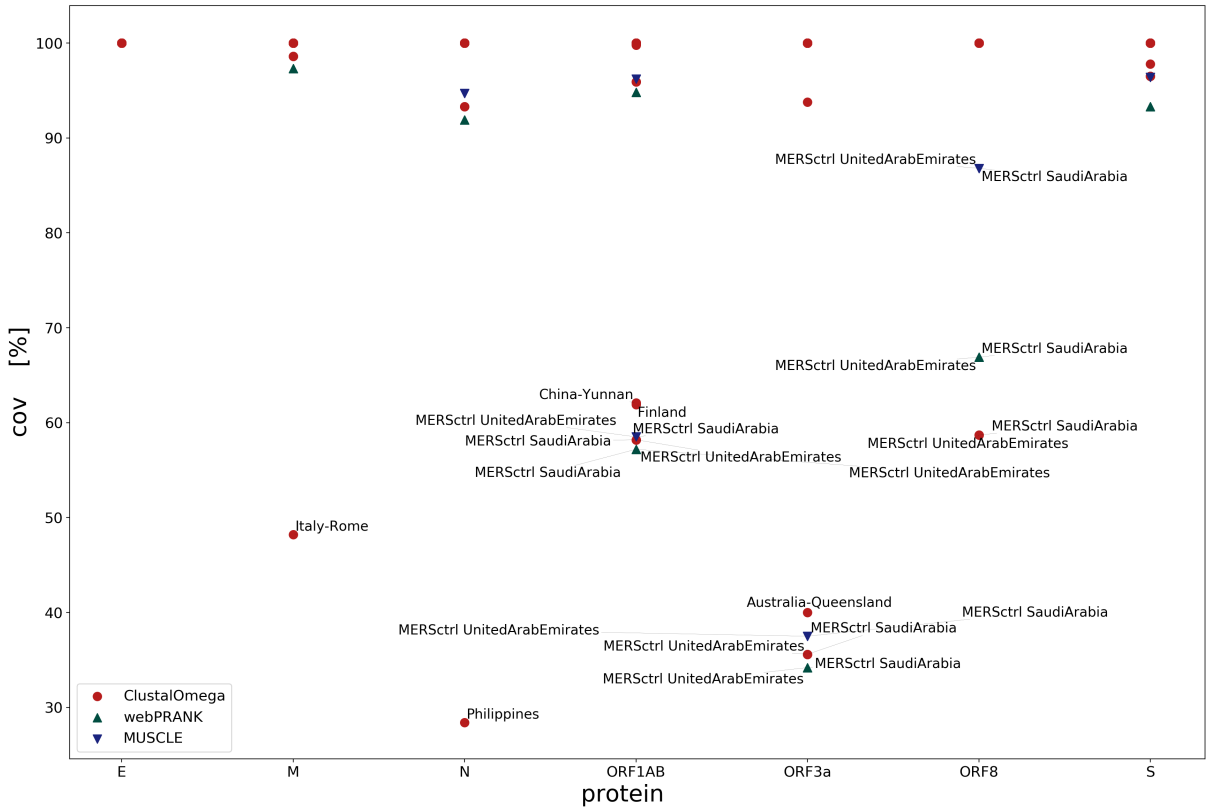

FIGURE 1

cov values calculated with MView for all sequences of each protein sequence, using China's Dec2019 sequence as reference with Equation (1).

As detailed in Figure 1, the MERS sequences have a low percentage coverage in relation to the SARS-CoV-2 sequences with in about60% values. Some other SARS-CoV-2 sequences share low percentage values with the MERS sequences, following are some explanations:

- The EBI's database stores a short sequence of M_Italy-Rome_Jan2020, for that reason the alignment added a sequence of gaps in order to match the alignment. M_Italy-Rome_Jan2020 covers a small fragment of the hole sequence, hence the small cover-

age value. Similar cases apply for N_Philippines_2020Jan23, N_Philippines_2020Jan26, N_Philippines_2020Feb06, and ORF3a_Australia-Queensland_Feb2020

- ORF1a_Finland_29Jan2020, and ORF1a_China-Yunnan_17Jan2020 were expected to have a low coverage as the reference sequence assess protein ORF1ab (not ORF1a). They were added to confirm the difference and discard a possible typo mistake.

In summary, the cov parameter describes how many 'gaps' are added to the alignment to achieve a better fit.

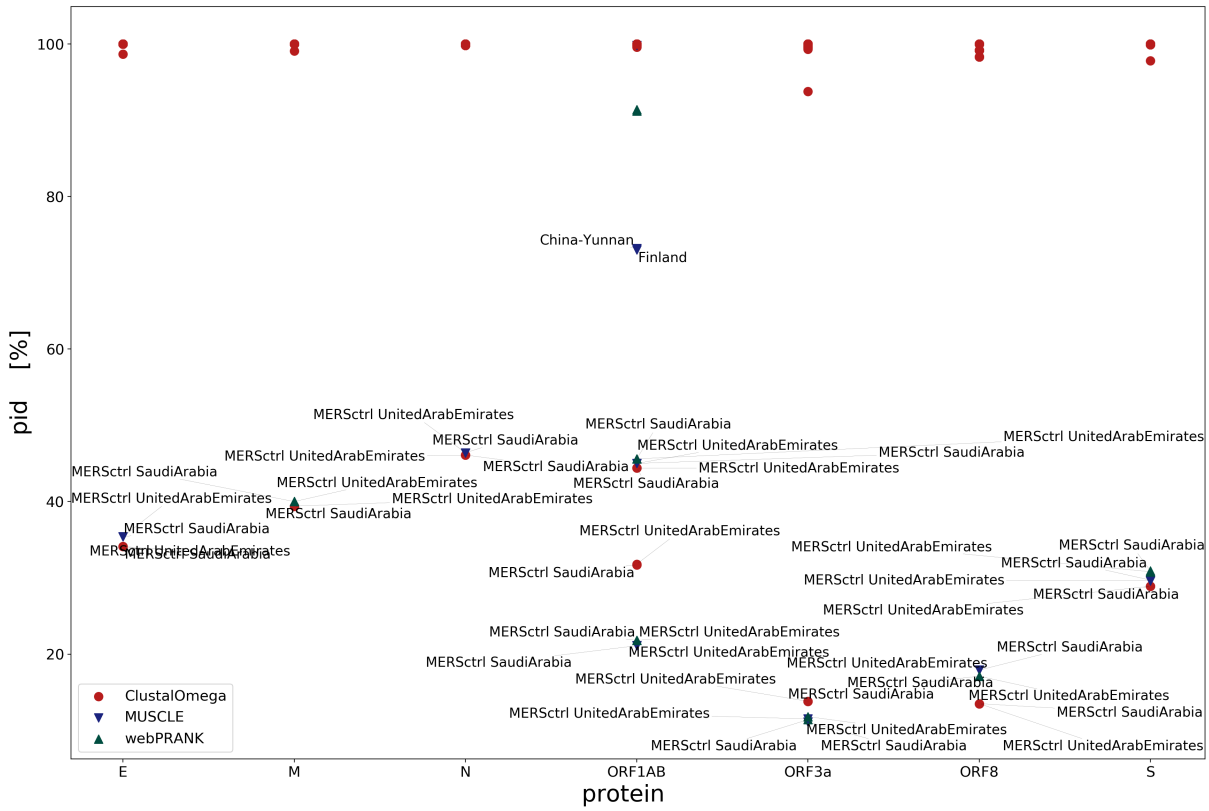

FIGURE 2

pid values calculated with MView for all sequences of each protein sequence, using China's Dec2019 sequence as reference with Equation (2).

Figure 2 plots the pid calculations. As expected, only the MERS sequences have low identification values (confirming that they are different from the SARS-CoV-2). ORF1a_Finland_29Jan2020, and ORF1a_China-Yunnan_17Jan2020 also share a low pid, further explaining Figure 1's conclusion.

References

[1] Mark J.G. Bakkers, Yifei Lang, Louris J. Feitsma, Ruben J.G. Hulswit, Stefanie A.H. de Poot, Arno L.W. van Vliet, Irina Margine, Jolanda D.F. de Groot-Mijnes, Frank J.M. van Kuppeveld, Martijn A. Langereis, Eric G. Huizinga, and Raoul J. de Groot. Betacoronavirus Adaptation to Humans Involved Progressive Loss of Hemagglutinin-Esterase Lectin Activity. *Cell Host {&} Microbe*, 21(3):356–366, mar 2017. 10.1016/j.chom.2017.02.008

[2] CDC. Coronavirus | Human Coronavirus Types, Centers for Disease Control and Prevention, 2020. https://www.cdc.gov/coronavirus/types.html

Author biography

Antonio Osamu Katagiri Tanaka .

MNT16

A01212611