Semi-covariance coefficient analysis of spike proteins from SARS-CoV-2 and other coronaviruses for viral evolution and characteristics associated with fatality

Jun Steed Huang [1, 2], Rebecca Spencer [3, 4], Wandong Zhang [3, 4]

[1]School of Information Technology, Carleton University, Ottawa, Canada.

[2]Dept. of Computer Science, Jiangsu University, Suqian, Jiangsu, China

[3] Human Health Therapeutics Research Centre, National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6

[4] Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

Short title: Analysis of coronaviral spike proteins for viral characteristics

Corresponding:
Dr. Wandong Zhang:
1200 Montreal Road, Building M54,
Ottawa, Ontario
Canada K1A0R6
Tel: 1-613-993-5988
Emails: Wandong.Zhang@nrc-cnrc.gc.ca; wzhan2@uottawa.ca;

Dr. Jun (Steed) Huang:
Carleton University
Ottawa, Ontario
Canada
Email: Jun.Huang@carleton.ca; steedhuang@ujs.edu.cn

 doi:10.20944/preprints202103.0055.v1

Abstract

Complex modeling has received significant attention in recent years and is increasingly used to explain the statistical phenomenon with increasing and decreasing fluctuations such as the similarity or difference of spike protein charge patterns of coronaviruses. Different from the existing covariance or correlation coefficient methods in traditional integer dimension construction, this study proposes a simplified novel fractional dimension derivation with the exact Excel tool algorithm. It involves the fractional center moment extension to covariance, which ends up a complex covariance coefficient that is better than the Pearson correlation coefficient, in the sense that the nonlinearity relationship can be further depicted. The spike protein sequences of coronaviruses were obtained from the GenBank and GISAID database, including the coronaviruses from pangolin, bat, canine, swine (three variants), feline, tiger, SARS-CoV-1, MERS, and SARS-CoV-2 (including the strains of Wuhan, Beijing, New York, German, and UK variant B.1.1.7) were used as the representative examples in this study. By examining the values above and below the average/mean based on the positive and negative charge patterns of the amino acid residues of the spike proteins from coronaviruses, the proposed algorithm provides deep insights into the nonlinear evolving trends of spike proteins for understanding the viral evolution and identifying the protein characteristics associated with viral fatality. The calculation results demonstrate that the complex covariance coefficient analyzed by this algorithm is capable of distinguishing the subtle nonlinear differences in the spike protein charge patterns with reference to Wuhan strain SARS-CoV-2 for which the Pearson correlation coefficient may overlook. Our analysis reveals the unique convergent (positive correlative) to divergent (negative correlative) domain center positions of each virus. The convergent or conserved region may be critical to the viral stability or viability; while the divergent region is highly variable between coronaviruses suggesting high frequency of mutations in this region. The analyses show that the conserved center region of SARS-CoV-1 spike protein is located at amino acid residues 900, but shifted to the amino acid residues 700 in MERS spike protein, and then to amino acid residues 600 in SARS-COV-2 spike protein, indicating the evolvement of the coronaviruses. Interestingly, the conserved center region of the spike protein in SARS-COV-2 variant B.1.1.7 shifted back to amino acid residues 700, suggesting this variant is more virulent than the original SARS-COV-2 strain. Another important characteristic our study reveals is that the distance between the divergent mean and the maximal divergent point in each of the viruses (MERS>SARS-CoV-1>SARS-CoV-2) is proportional to viral fatality rate. This algorithm may help to understand and analyze the

 doi:10.20944/preprints202103.0055.v1

evolving trends and critical characteristics of SARS-COV-2 variants, other coronaviral proteins and viruses.

Key words: Fractional complex moment, SARS-CoV-2, coronaviruses, spike protein sequence, Pearson correlation coefficient, semi-covariance coefficient, positive-correlative and negative-correlative domains

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

1. Introduction

Complex algorithms are used to analyze real-world implementations, i.e., it comes as the trusted analytic solution, but typically tends to have challenges in software implementation complexity requiring simpler software solution by using complex theory. Complex algorithms have received significant attention in recent years and are increasingly used to solve real-world problems among which are the combination of two or more algorithms involving numerical algorithms, analytic calculation [1], and other computational techniques, such as artificial intelligence [2-4], gene analysis systems [3] or gene simulation [4].

The Fractal DNA hypothesis (FDH) was first introduced by at least three groups independently in 1992 [5]. Being different from the traditional DNA hypothesis, RNA and protein analysis is based on fragment length between the domains with electrical charges. Especially, it emphasizes the influences on the behaviors of charges caused by the difference of information reception and lengths of expression or neighbor status observing the existence of fractal structure in stable DNAs [6]. Some work around FDH on RNA has been reported recently [7] where the genetic sequences were converted to binary numbers, purines converted to -1 and pyrimidines converted +1. The dimension order is found to be SARS-CoV-1 > SARS-CoV-2 > MERS, which differs from the time evolution order. Thus, we wish to further examine the similar relationship among the coronaviruses. SARS-CoV-2 virus is among the longest positive single-stranded RNA virus and its protein folding/tertiary structure is closely related to the charges of the amino acid residues. It is therefore important to examine the charging structure/patterns or nonlinear correlation pattern of the spike protein of SARS-CoV-2 as compared to the spike proteins from other coronaviruses to understand viral evolution and the characteristics of spike proteins associated with viral fatality.

The FDH does not distinguish the values above or below the mean (average) of the DNA fragment length between the gene signatures. Our algorithm used in this study focuses on distinguishing the values above and below the mean (average) to calculate the semi-covariance (based on the original semi-variance principle [27] for financial risk analysis initiated by 1990 Nobel Prize winner Harry M. Markowitz) coefficient of the spike protein sequences from coronaviruses, including SARS-CoV2. The higher value above the mean indicates higher similarity and increased evolutionary conservation, while the lower value below the mean indicates more dis-similarity and increased mutations. Analysis with our algorithm can be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

carried out rapidly by running Microsoft Excel sheet tool. In our study, the traditional Pearson correlation coefficient for the spike protein sequences of coronaviruses [8] is also calculated for comparison [9]. By imaging the charge similarity covariance as a weight (gravity or Coulomb force) on the rod of the axis, the weight center of semi-covariance coefficient is calculated to examine the evolving weight center [both convergent (positive correlative) center/region and divergent (negative correlative) center/region] shifting pattern of the spike protein sequences of coronaviruses [10] and identify spike proteins characteristics associated with viral fatality.

2. Materials and Methods

2.1. Coronaviruses and spike protein sequences

The coronavirus spike protein sequences used in this study were obtained from the NCBI GenBank and the GISAID database, including SARS-CoV-2 (the sequences of the virus strains isolated in Wuhan, Beijing Xinfadi wholesale market, Germany, New York, UK Wales and New York Zoo tiger), SARS-CoV-1, Middle East respiratory syndrome (MERS), bat coronavirus (RaTG13), pangolin coronavirus, feline coronavirus, canine coronavirus, and swine coronaviruses [Swine Transmissible gastroenteritis virus (Swine-stomach), swine enteric coronavirus (Swine-Ent), and porcine respiratory coronavirus (Swine-Res)]. The sequence ID from the GenBank and GISAID database are listed in Table 1.

2.2. Hypo, Hyper or Gauss Variances and Covariance

The complex parameter is a measure of fluctuation-term memory time series. It relates to the autocorrelation time series, and the derivatives of Laplace transformation (Frequency Spectrum) time series or the momentum generation function at the origin [11]. Studies involving the complex parameter were originally developed by Jerome Cardan (1501-1576) for solving algebra equations.

In order to calculate the Fractional version of the center momentum, we need to first generalize the Binomial formula from integer domain to real domain, as described previously [12]. Here we only show the formula:

$$\mu_k = E(\xi - E)^k = E\left(\sum_{i=0}^{\infty} \left(\frac{k}{i}\right)(-1)^i \xi^{k-i}(E\xi)^i\right) \#(1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

$$= \sum_{i=0}^{\infty} \left(\frac{k}{i}\right)(-1)^i E(\xi^{i-k}) E^i \xi \qquad (1 < k < 3) \#(2)$$

When k=2, it is Gauss variance; k<2 is hypo version; k>2 is hyper version. Factorial of fractional k is calculated by Gamma function seen below:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \#(3)$$

From which we can also have the covariance counterpart:

$$v_{2k} = E(\xi - E\xi)^k (\eta - E\eta)^k \#(4)$$

$$= E(\sum_{i=1}^{\infty} \left(\frac{k}{i}\right)(-1)^i \xi^{k-i} (E\xi)^i \sum_{i=1}^{\infty} \left(\frac{k}{i}\right)(-1)^i \eta^{k-i} (E\eta)^i) \#(5)$$

$$\rho = \frac{v}{\mu_\xi \times \mu_\eta} \#(6)$$

In sum, the product of two differences (the two values above the mean and the two values below the mean) on the same side of the mean value will be the real part, or the convergent part so that we can call it positive correlation covariance coefficient. The product that is on the opposite side of the mean will be the imaginary part, or the divergent part so that we can call it negative correlation covariance coefficient [13].

3. Results

3.1. Excel calculations of semi-covariance for spike proteins from SARS-CoV-2 and other coronaviruses

To compare and prove the usefulness of the simplified complex variances, we compare the correlation of SARS-CoV-2 viral spike protein sequence with other coronavirus spike protein sequences [14]. Since Excel is not capable of handling the imaginary number, we simplify the calculation with integer power, but separate the positive and negative covariance signs [15]. Because coronaviral spike proteins have different electrical charge levels [16], we normalize the covariance by the variance respectively just as the Pearson calculation does [17]. We calculated the sequences of the spike proteins [18] and plotted the curve starting from the N-terminus to the C-terminus. By using the moving window (a typical peptide) of 16 neighborhood amino acid residues [19], we calculated the covariance and average/mean over the same period of sequences to make the curve visually smooth for easier comparisons [20].

We defined the conserved centers or regions of spike proteins for different coronaviruses. For example, the conserved centers of SARS-CoV-1, pangolin and bat coronaviruses are located at

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

the amino acid residues 905, 906 and 904, respectively. These viral spike proteins are set as the 900 series (Table 1). The conserved center is defined as the weight center of the spike protein sequence at which the charge pattern before and after those points is the same. The conserved center for SARS-CoV-2 is at the amino acid residue 658, and the conserved center for feline coronavirus is at the amino acid residue 683 (Table 1). SARS-CoV-2 and feline coronaviral spike proteins are defined as the 600 series. The conserved centers for MERS and three swine coronaviruses range from amino acid residues 727 to 784, hence defined as the 700 series. So is the latest SARS-CoV-2 variant B117 from UK, the conserved center is at 702, which suggests that the charge pattern of variant B117 spike protein has evolved to 700 series from 600 series and that the variant B117 may be more virulent than the original SARS-COV-2 strain.

Figures 1-5 are the calculation results from our algorithm of semi-covariance coefficient for spike protein of Wuhan SARS-CoV-2 in comparison with spike proteins of other coronaviruses listed in Table 1. The spike protein sequences were analyzed with index order from animal coronaviruses (pangolin, bat, canine, swine, feline, and tiger) and human coronaviruses (SARS-CoV-1, MERS, and SARS-CoV-2) [21]. Figure 1 presents the calculation results of the spike proteins for human coronaviruses including 600 (SARS-CoV-2)//700 (MERS)//900 (SARS-CoV-1) series of spike proteins semi-covariance selected based on Table 1. Figure 2 presents the analysis for coronaviruses whose conserved center is located from the spike protein amino acid residues 900 to 999 (900 series) as described above. Figure 3 is for 800 and 700 series. Figure 4 is for 700 and 600 series. Figure 5 is for 600 series. Figure 6 is for 600 and 700 series.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

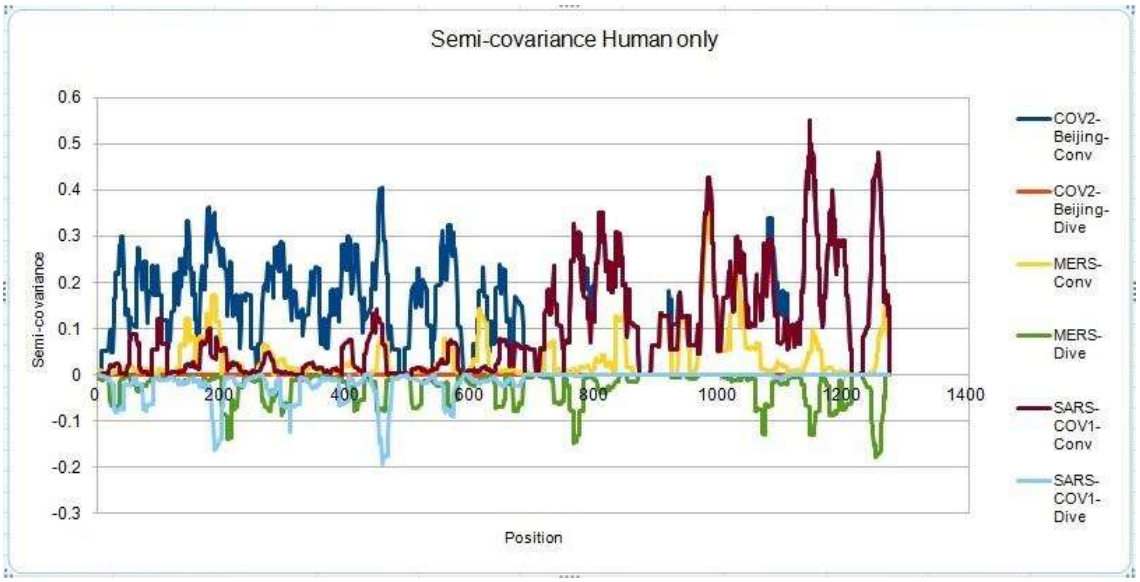

Fig. 1. Semi-covariance coefficient among the spike proteins from human coronaviruses SARS-CoV-2, SARS-CoV-1 and MERS: the spike protein sequence of Wuhan SARS-CoV-2 is used as a template for comparison with the spike proteins of SARS-CoV-2 viruses isolated from Beijing Xinfadi wholesale market (carrying D614G mutation), SARS-CoV-1, and MERS. Mathematically speaking, the diagram/curve above the X-axis is the positive correlation (convergent or conv in the figure). The higher the value, the greater the similarity of charge patterns between the compared viral spike proteins. SARS-CoV2 isolated in Beijing Xinfadi wholesale market is the same as the SARS-CoV-2 isolated in Wuhan throughout the entire sequence except D614G mutation. SARS-CoV-1 shows a similar pattern with SARS-CoV-2 after amino acid residue 700; while MERS shows a similar pattern with SARS-CoV-2 only around the amino acid sequence 1000. The 2nd similar region of MERS spike protein sequence with SARS-CoV-2 lies around amino acid residue 200. The diagram/curve below the X-axis is the negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged, thus the greater the dis-similarity between the compared viruses. MERS has more opposite charges around amino acid position 200, 800 and 1200 as compared to SARS-CoV-2, SARS-CoV-1 has a few opposite charges around amino acid position 200 and 450 as compared to SARS-CoV-2, indicating more similarity between the spike proteins from SARS-CoV-1 and SARS-CoV-2 but not between SARS-CoV-2 and MERS.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

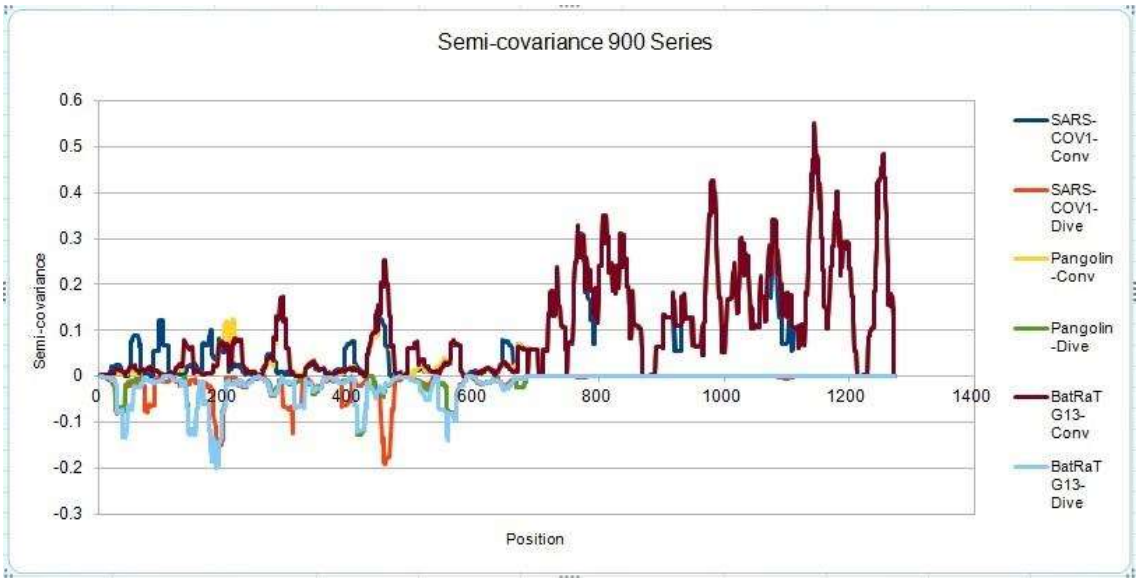

Fig. 2. Semi-covariance coefficient of SARS-CoV-2 spike protein with the spike proteins from SARS-CoV-1, pangolin and bat coronaviruses (900 series): The diagram/curve above the X-axis is positive correlation (convergent or conv in the figure). The higher the value, the greater the similarity of the charge patterns between the compared viruses. The spike protein of SARS-CoV-1 is similar to the spike protein of Wuhan SARS-CoV-2 from amino acid residues 700 onwards, as well as the spike proteins from pangolin and bat coronavirus. The spike proteins of SARS-CoV-1, bat and pangolin coronaviruses overlap each other more after amino acid residues 700. The diagram/curve below the X-axis is negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged, thus the more dis-similarity between the compared viruses. The spike proteins of the pangolin and the bat peak around position 200, suggesting that the charge pattern is not similar at this region or the mutations occurred more at this region between SARS-CoV-2 and the pangolin/bat coronaviruses. Similarly, SARS-CoV-1 peaks around 200 and 450 amino acid residue positions, suggesting that the charge patterns are different between SARS-CoV-1 and SARS-CoV-2 at this region or the mutations have made this region different between the two viruses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

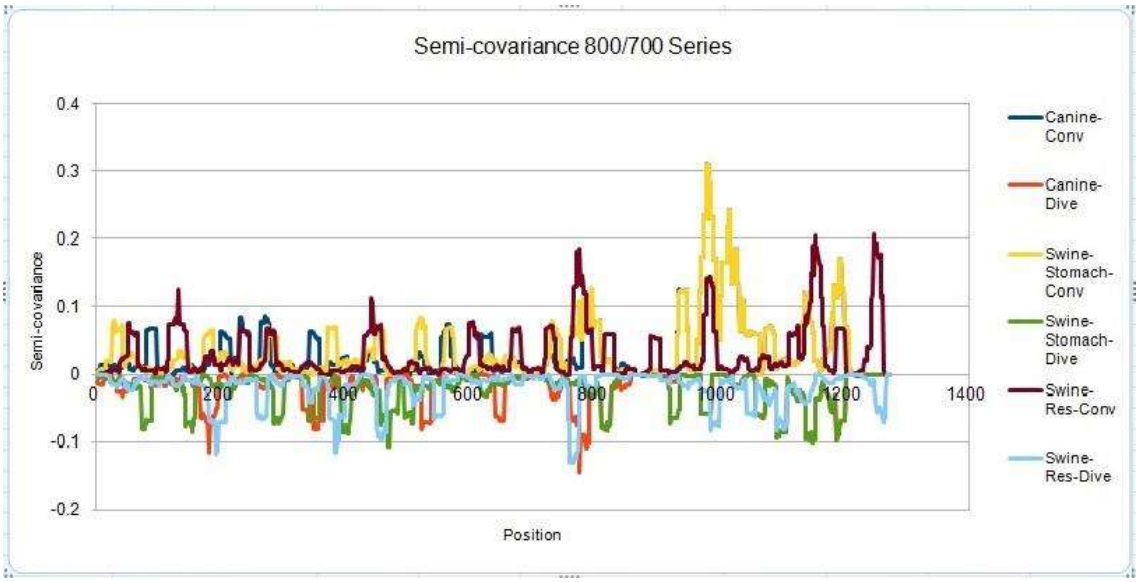

Fig. 3. Semi-covariance coefficient of SARS-CoV-2 spike protein with the spike proteins from MERS, canine and swine coronaviruses (700/800 series): The diagram/curve above the X-axis is the positive correlation (convergent or conv in the figure) between the compared viruses. The higher the value, the greater the similarity of charge patterns among the compared viruses. Canine and swine transmissible gastroenteritis virus (Swine-Stomach) are similar to Wuhan SARS-CoV-2 from the amino acid residue 1000 onwards, as well as porcine respiratory coronavirus (Swine-Res), but with less similarity. The diagram/curve below the X-axis is the negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged and the greater the dis-similarity between the compared viruses are. The spike proteins from canine, swine transmissible gastroenteritis virus (Swine-Stomach) and Swine-Res coronaviruses have opposite charge patterns at different positions. The spike proteins from canine and Swine-Res coronaviruses peak around 800, and the one from Swine-Stomach coronavirus peaks around amino acid residue positions 450 and 1150.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

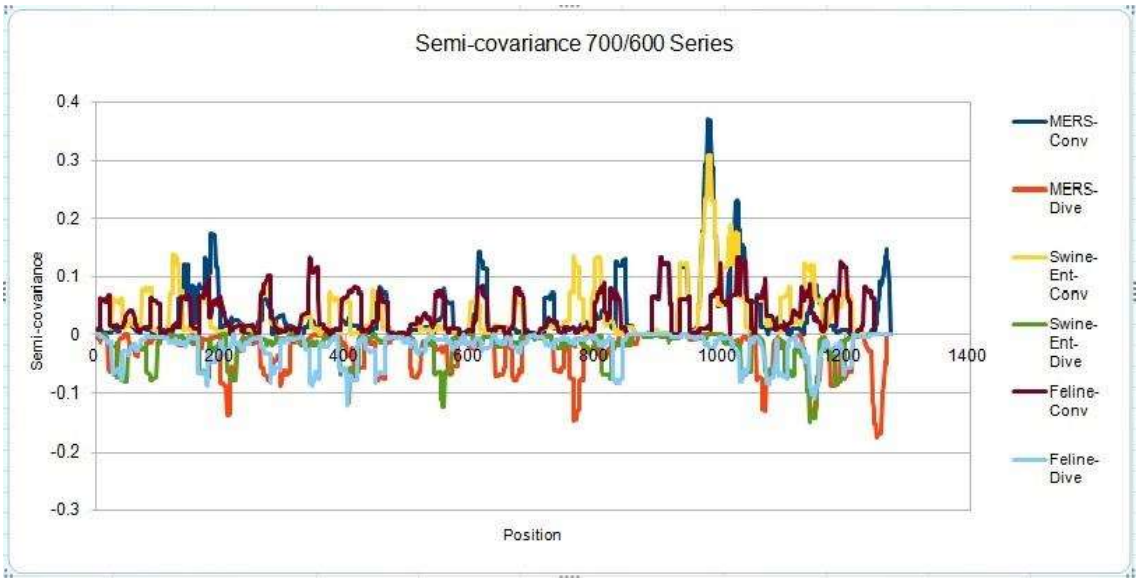

Fig. 4. Semi-covariance coefficient of SARS-CoV-2 spike protein with the spike proteins from MERS, swine and feline coronaviruses (700/600 series): The diagram/curve above the X-axis is the positive correlation (convergent or conv in the figure). The higher the value, the greater the similarity of charge patterns between the compared viruses. The spike proteins from MERS and Swine-Ent coronaviruses are similar to Wuhan SARS-CoV-2 around amino acid residues 1000; while the spike protein from feline coronavirus is not similar to Wuhan SARS-CoV-2. The diagram/curve below the X-axis is the negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged, the greater the dis-similarity between the compared viruses. The spike proteins from MERS, Swine-Ent and feline coronaviruses have opposite charge patterns at different positions. The spike proteins of feline and Swine-Ent coronaviruses peak around 400 and the spike protein of MERS peaks around amino acid residue positions 750 and 1250.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

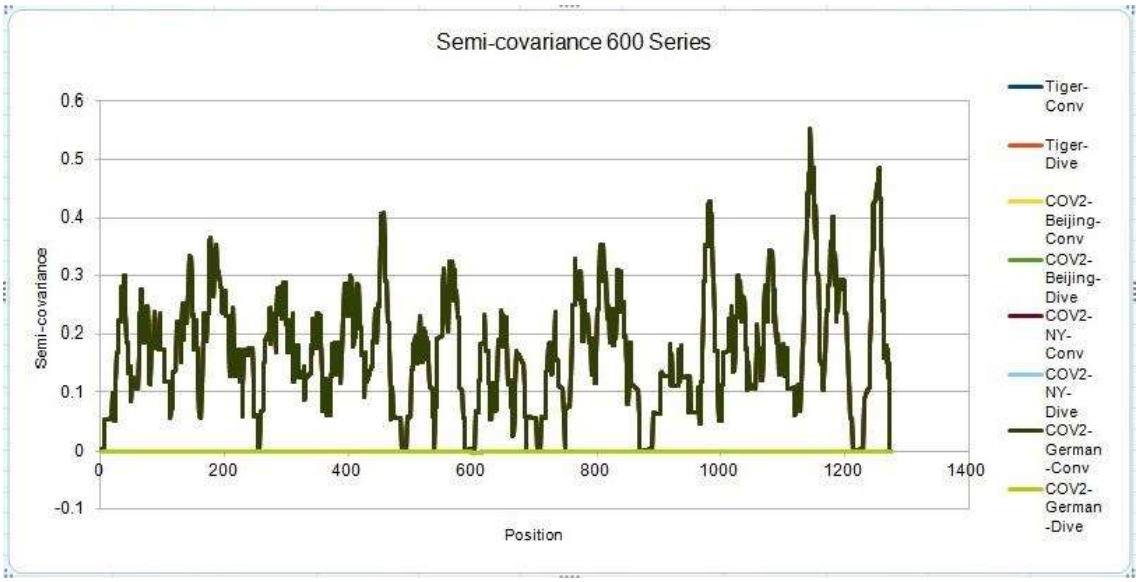

Fig. 5. Semi-covariance coefficient of Wuhan SARS-CoV-2 spike protein with the spike proteins of SARS-CoV-2 isolated in Beijing, New York, Germany, and New York Zoo tiger (600 series): The diagram/curve above the X-axis is the positive correlation (convergent or conv in the figure). The higher the value, the greater the similarity of charge patterns between the compared viruses. SARS-CoV-2 isolated in Beijing Xinfadi wholesale market, New York, Germany and the New York zoo tiger overlap and are almost identical to Wuhan SARS-CoV-2. The diagram/curve below the X-axis is the negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged and the greater the dis-similarity between the compared viruses are. SARS-CoV-2 spike proteins from Beijing, New York, Germany and the New York zoo tiger carry D614G mutation and have the only opposite charge at amino acid residue 614 position (D614G mutation as reported in the literature).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

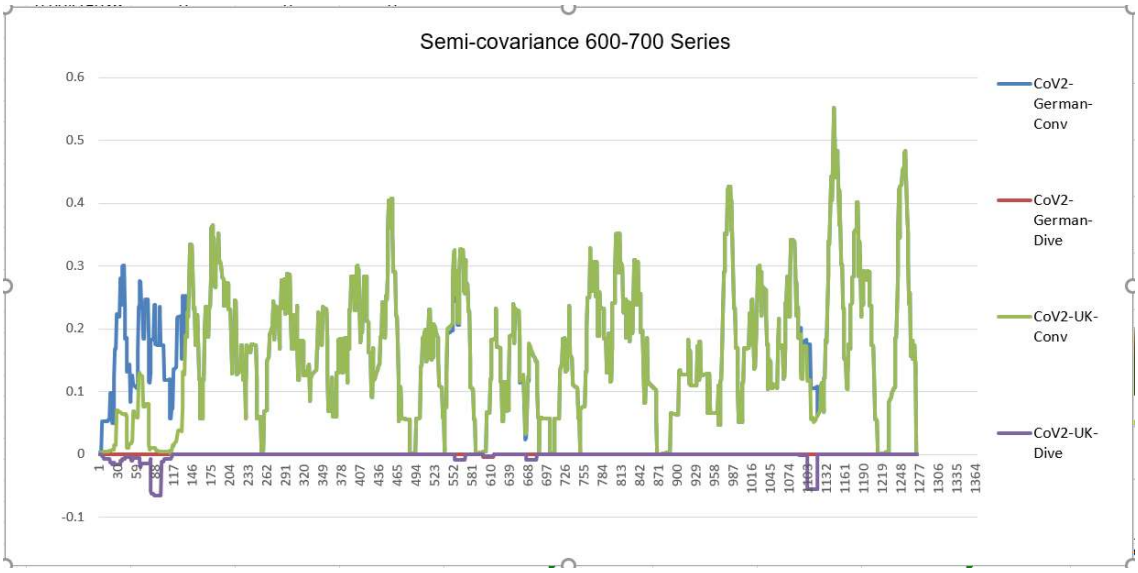

Fig. 6. Semi-covariance coefficient of Wuhan SARS-CoV-2 spike protein with the spike proteins of SARS-CoV-2 isolated in UK variant B117 (700 series) and Germany (600 series): The diagram/curve above the X-axis is the positive correlation (convergent or conv in the figure). The higher the value, the greater the similarity of charge patterns between the compared viruses. SARS-CoV-2 variants isolated in UK Wales (B117) and Germany are almost identical to Wuhan SARS-CoV-2, except the beginning part of UK is mutated back to SARS-COV-1. The diagram/curve below the X-axis is the negative correlation (divergent or dive in the figure). The lower the value, the more oppositely charged and the greater the dis-similarity between the compared viruses are. The UK variant B117 has not only opposite charge around 600 amino acid residue position (D614G mutation as reported in the literature) but also at the old one around 100 (SARS-CoV-1) and at a new position 1000. The mutation sites occur towards both sides of 600 series. It flips back more like bat coronavirus as well. The gaps between the mutation sites are as follows: 69=3x23; 73; 355=3x71; 69=3x23; 44=2x2x11; 67; 35=5x7; 266=2x7x19; 136=2x2x2x17; 155=5x31. There are three groups of prime numbers involved. The first group is 2,3,5, which belongs to cusps modular (Langlands) prime number. It might be related to the Fractal shell like growing structure. The second group is 7,11,19,23, which belongs to 4k+3 prime number, also called Gaussian prime number. The latter is a closed field number on a complex plane, meaning that the numbers form a total ordered chain. It might be attribute to the 3D chain structure of the spike protein. The third group is 31,67,71,73, which belongs to the prime numbers of binary digits prime number. It might be attributed to the long folding structure of the protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

3.2. Pearson and semi-covariance coefficient analysis of spike proteins from coronaviruses

Table 1 compares the Pearson correlation coefficient analysis with semi-covariance coefficient analysis for coronaviral spike proteins. From Table 1, it shows that the Pearson correlation coefficient only reflects the variation after the cancellation of up and down correlation [22]; however, our semi-covariance coefficient reflects the direction of the variations before the cancellation of correlation [23]. Table 2 uses the spike protein sequence from Wuhan strain SARS-CoV-2 as a template to compare with SARS-CoV-2 isolated in Beijing (carrying D614G mutation), SARS-CoV-1, and MERS. It incorporates fatality rates to identify critical amino acid regions associated with mortality. As compared to SARS-CoV-2, there are 74 amino acid residues in MERS spike protein sequence that are critical to MERS-associated fatality, and there are 18 amino acid residues that are associated with SARS-CoV-1 fatality. There are only 9 amino acid residues in the Beijing strain viral spike protein that are different from Wuhan strain SARS-CoV-2 spike protein. The correlation coefficient of the analysis for these critical amino acids in the spike proteins associated with fatality is R=0.9981 among the three coronaviruses infecting humans. The similar calculation for the ratio of Mutation Coulomb force center to maximum Coulomb force point leads to R=0.9958. Divergent Coulomb intensity dictates the fatality.

To visually examine the nonlinear correlation relationship between Wuhan strain SARS-CoV-2 and the rest of the coronaviruses, we further plotted scatter graphs where the X-axis is the charge variation over 16 neighborhood amino average of Wuhan strain SARS-CoV-2, and the Y-axis is the charge variation over 16 neighborhood amino average of the respective virus (Fig. 7-10). As it can be seen that some of them have nonlinear relationships. Pearson correlation may not be good enough to depict all of them. The second quadrant and fourth quadrant represent the strong mutation part where the charge is reversed, the first quadrant and the third quadrant are the weak mutation part where the charge is not reversed (Fig. 7-10).

It can be seen that there are a combination of linear and nonlinear relationships in Fig. 7A and 7B; while Fig. 7C shows a linear relationship. Fig. 8A, 8B and 8C show a linear relationship for the scatter patterns of the spike proteins from the viral strains isolated in German, New York and Beijing relative to that of Wuhan strain, indicating high similarity. Fig. 7C is identical to those of Fig. 8A-8C, suggesting that the same strain of the SARS-CoV-2 was transmitted from human to the New York Zoo tiger. All of these viral strains carry the D614G mutation in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

spike protein. However, Fig. 8D shows the UK variant B.1.1.7 vs. Wuhan strain and there is a combination of linear and nonlinear relationships between them, indicating that the mutations in B.1.1.7 results in amino acid changes with opposite charges as compared to Wuhan strain. Figure 9A shows a combination of linear and nonlinear relationships between SARS-CoV-1 and Wuhan strain SARS-CoV-2 spike proteins; while Fig. 9B shows a nonlinear relationship between MERS and Wuhan strain SARS-CoV-2 spike proteins, indicating a strong dis-similarity between them. Figure 10 shows nonlinear relationships between those animal viral spike proteins and Wuhan strain SARS-CoV-2 spike protein, indicating a strong dis-similarity between them. However, the local piecewise similarity island pattern is still clearly seen. That means they are still related somehow.

4. Discussion

This study presents the construction of a complex covariance for fractional analysis of coronavirus spike proteins by using fractional moment based simple algorithm coded in Excel Sheet. The analysis with our novel complex model reveals additional performance index over the traditional real model, such as Pearson correlation coefficient. Our model compares the traditional Pearson calculation of the integer dimension against the fractional dimension. The complex calculation shows the differences among viral spike proteins where the traditional covariance definition and calculation may overlook. Our study reveals the unique convergent (positive correlative) to divergent (negative correlative) centers of each virus and the distance/length between the positive- and negative-correlative centers/regions (Table 1). Interestingly, we found that the distance between divergent center (mean) and the maximal divergent point is associated with viral fatality. As compared to SARS-CoV-2 strain isolated in Wuhan, the distance between the divergent center (mean) and the maximal divergent point is located at the amino acid residue 614 in the SARS-CoV-2 viral strains isolated in Beijing, Germany, New York and New York Zoo tiger. This suggests those viruses are essentially the same except at amino acid 614 [D614G mutation, aspartate (D) to glycine (G)] also reported in the literature [24, 25]. While the distance between the divergent center (mean) and the maximal divergent point in the spike protein of SARS-CoV-1 is from the amino acid residues 309 to 338 (Table 1), the distance between the divergent center (mean) and the maximal divergent point in the spike protein of MERS is from the amino acid residue 214 to 698 as compared to the spike protein of Wuhan strain SARS-CoV-2 (Table 1). It is evident that the fatality rate caused by the virus is highly related to the distance between the divergent center

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0055.v1

(mean Coulomb force) and the maximal divergent Coulomb (force) point (Table 2). The longer the distance the more mutations (Coulomb force) and the more deadly the virus is. This region of MERS spike protein occurs with high frequency of mutations and may be responsible for its high fatality.

From Table 1, it shows that our complex coefficient reveals more dependency and trends of each protein sequence evolution [26]. In the past, the viral spike protein's conserved center evolves from the amino acid residue 900 in SARS-CoV-1 down to 600 in SARS-CoV-2. The conserved region or convergent center may be critical to the viral stability or viability. The conserved center/region of the viral spike protein has been shifted from SARS-CoV-1 at the amino acid residue 900 to amino acid residue 700 in MERS spike protein, and then shifted to amino acid residue 600 in SARS-CoV-2. The charge pattern of the SARS-CoV-1 spike protein sequence around 900 is similar to that of the MERS spike protein around 700, and similar to that of the SARS-CoV-2 around amino acid residue 600. Interestingly, the convergent center of the UK variant B.1.1.7 is shifted from 600 in SARS-CoV-2 strains (Wuhan, German, New York, and Beijing strains) to 700 (Table 1). The convergent center of UK variant B.1.1.7 spike protein in 700 is similar to those of MERS and swine coronaviral spike proteins (Table 1), which may indicate a more lethality as compared to the SARS-CoV2 strains isolated in Wuhan, German, New York, and Beijing. Our analysis suggests that the conserved center/region may be essential for the biology and evolution of the coronaviruses. This conserved center/region may be shifted to a new location in SARS-COV-2 variants or other novel coronaviruses.

5. Conclusion

In this study, we have analyzed spike protein charge patterns of coronaviruses by using our algorithm of semi-covariance (nonlinear) coefficient as compared to Pearson (linear) correlation coefficient. The analysis reveals additional performance index over Pearson analysis, such as both positive- and negative-correlative centers/regions in the spike proteins. The analysis provides in-depth understanding for the nonlinear viral evolution pattern and identifies the protein Coulomb characteristics associated with viral fatality. It is envisioned that this complex number model is a good alternative for covariance analysis of coronaviral spike proteins. This type of analysis may go beyond asymmetrical fluctuations to help in developing high dimensional Fractal theory. However, the simplified calculation is easier for practical analysis and applications. The simplified Excel sheet calculation is very easy to use, accurate

and forward compatible with traditional Pearson model and calculations. The example code is available from the Excel file on the github server (https://github.com/steedhuang/covid-19-gene-convertor). Our future work will look in other viral proteins with the same methodology for viral evolution and the Coulomb characteristics that are associated with viral fatality. More attention will be paid on the relationship between positive charges to infectivity.

Acknowledgement

The work in Dr. Zhang's lab is supported by a team grant on the Rapid Research Response to COVID-19 Outbreak awarded from the Canadian Institute of Health Research (CIHR) and by funding from the National Research Council of Canada. Thanks go to Lishen Wang from Jiangsu University for writing Python code to covert sequences into charges. Thanks also go to Mei Huang from Ottawa Hospital COVID-19 patient unit for proof reading and editing the final version.

Author contributions:

Conceptualization: Jun Huang, Wandong Zhang

Data curation: Wandong Zhang, Rebecca Spencer

Data analysis and figures: Jun Huang

Data extraction and biological explanation: Wandong Zhang, Jun Huang

Writing – original draft: Jun Huang, Wandong Zhang

Writing - revise/review & edit: Wandong Zhang, Jun Huang, and Rebecca Spencer

Declaration of competing interest: The authors declare no conflict of interest.

Reference

1. Zunino L, Tabak BM, Figliola A, Pérez DG, Garavaglia M, Rosso OA. A multifractal approach for Spike Protein inefficiency. Physica A. 2008; 387: 6558–6566.
2. Brooks, Rodney. Elephants Don't Play Chess. Robotics and Autonomous Systems. 1990; 6: 3–15.
3. Gerla G. Fuzzy Logic Programming and fuzzy control. Studia Logica. 2005; 79: 231-254.
4. Davidovitch L, Parush A, Shtub A. Simulation-based Learning: The Learning-Forgetting-Relearning Process and Impact of Learning History, Computers & Education, Vol. 50, No. 3, 866–880, 2008.
5. Edgar E. Peters. Fractal DNA Analysis: Applying Chaos Theory to Investment and Economics, Wiley, 1994.
6. Xu Q, Bai Y. Semiparametric statistical inferences for longitudinal data with nonparametric covariance modelling. Stats A Journal of Theoretical & Applied Stats. 2017; 51(6): 1-24.
7. de Salazar e Fernandes, T., de Oliveira Filho, J.S. & da Silva Lopes, I.M.S. Fractal signature of coronaviruses related to severe acute respiratory syndrome. *Res. Biomed. Eng.* (2020). https://doi.org/10.1007/s42600-020-00069-5
8. Liu CS, Chang MS, Wu X, Chui CM. Hedges or safe havens—revisit the role of gold and usd against gene: a multivariate extended skew-t copula approach. Quantitative Finance. 2016; 1-27.
9. Donald SG, Hsu YC, Barrett GF. Incorporating covariates in the measurement of welfare and inequality: methods and applications. Econometrics Journal. 2012; 15(1): C1-C30.
10. Gillard J. Large covariance and autocovariance matrices; patterned random matrices. Journal of the Royal Statistical Society Statistics in Society Series A. 2018; 182(PT.2), 714-714.
11. Torshin IY, Harrison RW, Weber IT, Petock JM. Identification of protein folding cores using charge center model of protein structure. Scientific World Journal. 2014; 2(11): 236-237.
12. Zou Q, Hu Y, Huang JS. Definition of Complex Hurst and Fractional Analysis for Spike Protein Fluctuation. In Gen M, Kim K, Huang X, Hiroshi Y. (eds) Industrial

Engineering, Management Science and Applications. Lecture Notes in Electrical Engineering, Vol. 349. Springer, Berlin, Heidelberg, 2015.

13. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, et al. Long-range correlations in nucleotide sequences. Nature. 1992; 356(6365): 168-170.
14. Car Z, Egota SB, Aneli N, Lorencin I, Mrzljak V. Modeling the spread of covid-19 infection using a multilayer perceptron. Computational and Mathematical Methods in Medicine. 2020; 2020: 1-10.
15. Li WT, Marr TG, Kaneko K. Understanding Long-Range Correlations in DNA-Sequences. Oji International Seminar on Complex Systems: from Complex Dynamical Systems to Sciences of Artificial Reality: from Complex Dynamical Systems to Sciences of Artificial Reality. Elsevier North-Holland, Inc. (1994)
16. Li W, Kaneko K. Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. Europhysics Letters (EPL). 1992; 17(7): 655-660.
17. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A. 1993; 90(15): 7176–7180.
18. Sabeena M, Hanan B, Musa G. Peptide-Protein Interaction Studies of Antimicrobial Peptides Targeting Middle East Respiratory Syndrome Coronavirus Spike Protein: An In Silico Approach. Adv Bioinformatics. 2019; 2019: 6815105.
19. Wentian Li. Generating Non-trivial Long-Range Correlations and 1/f Spectra by Replication and Mutation. International Journal of Bifurcation and Chaos. 1992; 2: 137-154.
20. Xu PR, Fu WJ, Zhu LX. Shrinkage estimation analysis of correlated binary data with a diverging number of parameters. Science China Mathematics. 2013; 56: 359-377.
21. Bianchi M, Benvenuto D, Giovanetti M, Angeletti S, Pascarella S. SARS-CoV-2 envelope and membrane proteins: structural differences linked to virus characteristics? BioMed Research International. 2020; 2020: 1-6.
22. Ayal BG, Noam A, Guihem F, Yuri IW, Feng Z, Eugene V. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci U S A. 2020; 117:15193-15199.
23. Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters. 1992; 68(25): 3805-3808.
24. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans

CM, Freeman TM, de Silva TI; Sheffield COVID-19 Genomics Group, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812-827.e19. doi: 10.1016/j.cell.2020.06.043. Epub 2020 Jul 3. PMID: 32697968; PMCID: PMC7332439.

25. Grubaugh ND, Hanage WP, Rasmussen AL. Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear. Cell. 2020;182(4):794-795. doi: 10.1016/j.cell.2020.06.040. Epub 2020 Jul 3. PMID: 32697970; PMCID: PMC7332445.
26. Tim S, Jennifer B, Aimee L, Tony P. COVID-19 Drug Therapy. Elsevier. July 15, 2020.
27. Harry Markowitz, Portfolio Selection, The Journal of Finance, Vol. 7, No. 1., pp. 77-91, Mar., 1952.