2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) | 978-1-6654-0373-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/UPCON50219.2020.9376568

# The SARS-CoV-2 and its Similarity to Other Bat-Derived SARS-like Coronaviruses: A Data-Driven Study

Mayank Sharma
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
mayanksharma8993@gmail.com

***Abstract*—In this paper, a data-driven comparison of the novel coronavirus strain (the SARS-CoV-2) with two bat-derived SARS-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21 is presented. Biological research has shown that the SARS-CoV-2 shows sequence identity most similar to these two bat-derived viruses. Through comparison of guanine-cytosine content and oligonucleotide composition, variations in their genomes were observed. These genomes were then translated into proteins and their amino acid concentrations were compared. Finally, various properties of the 4 important structural proteins (envelope, membrane, spike and nucleocapsid proteins) such as aromaticity, isoelectric point, instability index and amino acid count were compared for the three coronaviruses. Not only does this paper present useful insights on the new virus strain, but also on the properties that distinguish it from other similar coronaviruses.**

***Keywords*—Coronavirus, Genomics, Proteomics, Genomic Analysis, Bioinformatics**

## I. Introduction

The world has witnessed several different outbreaks of epidemics/pandemics in the past few decades. Examples include the Severe Acute Respiratory Syndrome (SARS) outbreak in 2002 that started in the Guangdong province of China [1] and the Middle East Respiratory Syndrome (MERS) outbreak of 2012 that was predominant in Middle-Eastern countries [2]. In the end of 2019, the city of Wuhan in the Hubei province of China experienced the emergence of a novel coronavirus disease, with most cases linked to a wet seafood market in Hunan. This novel coronavirus strain was initially called 2019-nCoV, but it has subsequently been renamed by the International Committee on Taxonomy of Viruses (ICTV). Now, the disease caused by the novel coronavirus strain, the SARS-CoV-2, is called COVID-19. Major symptoms of this disease include fever, cough, sputum production and shortness of breath [3]. Since the emergence of this disease in China, it has spread to almost every corner of the world. On the 11th of March 2020, World Health Organization (WHO) declared the disease as a pandemic. Up until the 8$^{th}$ of November 2020, about 49 million people have been infected from the disease across the globe, with a death toll crossing 1.24 million [4]. Worst-hit countries include the USA, India, Brazil, Russia and France.

Coronaviruses possess positive-sense single stranded [(+)ss] RNA genomes whose lengths extend from 26 to 32 kilobases [5]. Their virions have an average diameter size of 80-120 nm and a glycoprotein present at around 20 nm above the virion envelope gives them a crown (coronet) like appearance [6]. They belong to the family Coronaviridae and possess 4 subgroups - alpha, beta, gamma and delta coronaviruses. There are 4 structural proteins present in the virus - a spike (S) protein, an envelope (E) protein, a membrane (M) protein and a nucleocapsid (N) protein, among several other non-structural proteins.

Researchers have found out that the novel coronavirus strain, SARS-CoV-2, has sequence identity similar to two bat-derived SARS-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, as reported in [7-13]. Taking this finding into consideration, the aims of this paper are three-fold. First, genomes of the three coronaviruses are compared to find out differences in their GC (Guanine-Cytosine) content and oligonucleotide composition, and this is followed by global sequence alignment. Second, the genomes are translated into proteins and their amino acid constituents are compared. Finally, a comparative analysis of their structural proteins is presented along with sequence alignment results. The research presented in this paper includes computation done on genomic data obtained through NCBI's portal and not through biological wet-lab research.

## II. The SARS-CoV-2

Scientists and researchers across the globe have started studying the characteristics of the novel coronavirus strain, the SARS-CoV-2. It is a type of betacoronavirus [14] and is the 7th type of coronavirus known to infect human beings.

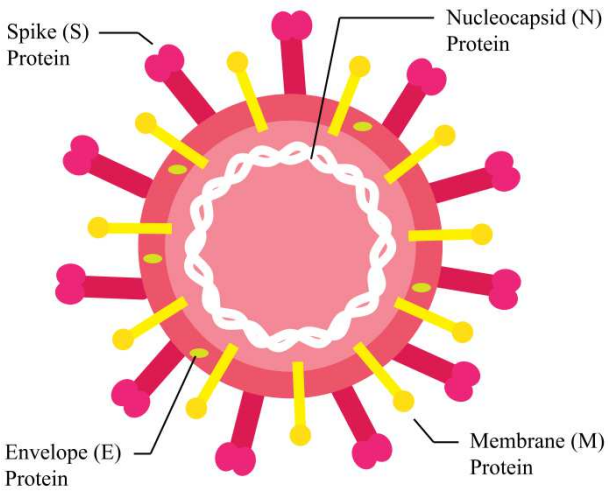


Fig. 1. Structural Proteins in the SARS-CoV-2

The SARS-CoV-2 genome has close to ~30000 nucleotides and encodes for four structural proteins, as shown by the majority of coronaviruses. These are the spike protein, envelope protein, membrane protein and nucleocapsid protein. Fig. 1 shows the structure of the SARS-CoV-2 and the structural proteins present in it. The virus enters the host by attaching itself to the host cell surface receptors through the S protein [15]. The E protein is a small protein that plays an important role in membrane permeability of the host cell and virus-host cell interaction [16]. The M protein is thought of as the central organizer for the coronavirus body. The N protein coats the positive sense single-stranded RNA genome of the virus and plays a vital role in replication and transmission of the virus [17].

## III. Data Used

Genomes are a sequence of the four nucleotide bases - adenine (A), guanine (G), cytosine (C) and thymine (T) - that contain the entire genetic information of an organism. Using various bioinformatics tools, these genome sequences can be read and manipulated in the form of *strings* of the 4 English alphabets - A, G, C and T, which are the 4 nucleotide bases that make up the genome. For the purpose of this study, fasta and genbank files of the genomes of SARS-CoV-2, bat-SL-CoVZC45 and bat-SL-CoVZXC21 were extracted from NCBI's database [18]. The NCBI accession numbers of these files and the number of nucleotides present in the genomes are presented in Table I.

TABLE I. NCBI Accession Number And Nucleotide Composition Of The Virus Strains

| NCBI Accession Number | Virus Strain | Number of Nucleotides |
|---|---|---|
| NC045512.2 | SARS-CoV-2 | 29903 |
| MG772934.1 | bat-SL-CoVZXC21 | 29732 |
| MG772933.1 | bat-SL-CoVZC45 | 29802 |

## IV. Comparison Of Genomes

### A. GC Content

Defined by equation (1), GC content is the percentage composition of G and C bases in a genetic material (DNA, RNA or genome). The G and C pairs are linked to each other through 3 hydrogen bonds while the A and T pairs are linked through 2 hydrogen bonds. Because of this reason, the GC content helps in determining the stability of sequences (higher is the GC content, higher is the stability of the genomic sequence). This information can be used by researchers in designing drugs that target specific parts of the sequence and destabilize them.

$$GC\ Content = \frac{n(G) + n(C)}{n(G) + n(C) + n(T) + n(A)} \quad (1)$$

The values of GC content for the 3 genomes is shown in Table II. It can be seen that the bat-derived coronaviruses have higher values of GC content as compared to SARS-CoV-2. These values are not drastically different because the genomic sequences are similar in identity but it can be concluded that the bat-derived coronavirus genomes are more stable as compared to the SARS-CoV-2 genome.

TABLE II. GC Content In The Genomes

| Virus Strain | GC Content |
|---|---|
| SARS-CoV-2 | 37.97% |
| bat-SL-CoVZXC21 | 38.82% |
| bat-SL-CoVZC45 | 38.90% |

### B. Oligonucleotide Composition

An oligonucleotide is defined as a short sequence of a DNA/RNA or a genome. Longer sequences are generally divided into shorter oligonucleotide sequences to study them in detail. Oligonucleotides are also referred to as "k-mers", where the alphabet 'k' refers to the size of the oligonucleotide. For example, a trimer (or a 3-mer) would be a sequence of the size 3 obtained from a larger genomic sequence. Since genomes are composed of 4 nucleotide bases, the total number of dimers/dinucleotides that can be possibly obtained from them are 16.

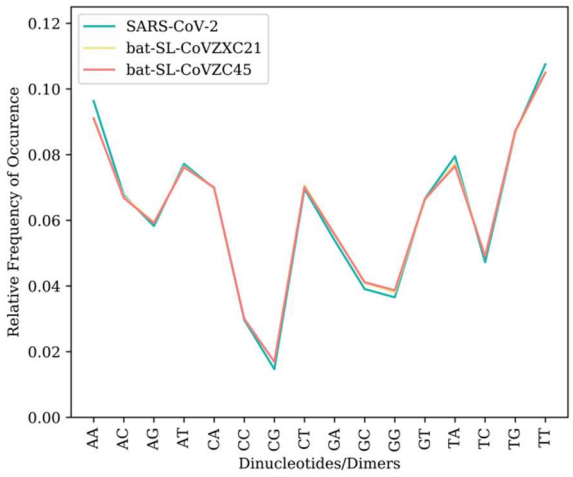


Fig. 2. Relative Frequency of Occurrence of Dimers in the Three Genomes

The relative frequency of occurrence of all these dimers in the 3 genome sequences is shown in Fig. 2. The line graph overlaps for both bat-derived coronaviruses, indicating little difference in the frequencies of occurrence of dimers. There are, however, little differences in frequencies of occurrence when compared to the SARS-CoV-2, particularly for the dimers AA, GA, GG, GC, CG and TA.

### C. Sequence Alignment

Sequence alignment is a technique that is used to find out sequence identity between genomes, and thus possibly establishing the existence of similar evolutionary, functional and structural relationships among them. In pairwise sequence alignment, similarity between two query sequences is found out. Pairwise alignment can further be divided into global and local sequence alignment. In a local sequence alignment, local regions with the highest similarity are found out, while in a global sequence alignment, the entire sequences are compared end to end.

One of the first and most common global sequence alignment algorithms was the Needleman-Wunsch Algorithm, which was proposed in 1970 [19]. However, the algorithm is not optimized for large sequences and hence takes a lot of memory space (precisely $O(MN)$ where $M$ and

$N$ are the size of the sequences to be aligned). In [20], an optimal algorithm was formulated that reduced the space to $O(N)$ where $N$ is the size of the sequence that is smaller out of the two being aligned. The EMBOSS Stretcher Tool [21], which works on this optimized Needleman-Wunsch Algorithm was used in this study to align the genome sequences. For DNA sequences, a gap open penalty of 16 and gap extend penalty of 4 is used by [21]. Similarly, for protein sequences, a gap open penalty of 12 and gap extend penalty of 2 is used.

Using the scoring criteria described above, the values of Percent Sequence Identity (PID) between SARS-CoV-2 & bat-SL-CoVZC45 and SARS-CoV-2 & bat-SL-CoVZXC21 were found out. It is obtained by dividing the total alignment score by the length of the shorter sequence, arithmetic mean of the length of the two sequences or other suitable metrics [22]. The PID values obtained are indicated in Table III. Results indicate very similar percentages to the ones listed in [7-13].

TABLE III. PERCENT SEQUENCE IDENTITY BETWEEN GENOMES

| Pair Aligned | PID Value |
|---|---|
| SARS-CoV-2/ bat-SL-CoVZXC21 | 87.5% |
| SARS-CoV-2/ bat-SL-CoVZC45 | 87.8% |

## V. TRANSLATION OF GENOMES INTO PROTEINS

Next, fasta files of the 3 genomes were translated into proteins using Biopython [23]. For positive sense single stranded RNAs, translation into proteins is a two-step process, described as follows:

### *A. Transcription of DNA into messenger RNA:*

In this process, Thymine (T) bases in the DNA are transcribed into Uracil (U) bases of the mRNA. For example, a DNA sequence 'ATTGC' is transcribed into an mRNA sequence 'AUUGC'.

### *B. Translation of messenger RNA into proteins:*

This is achieved by breaking down the genomic sequence into a series of codons. A codon is a sequence of 3 nucleotides in the mRNA, which codes for a specific amino acid or a stop signal. For example, the codon 'CUU' corresponds to the amino acid Leucine (L).

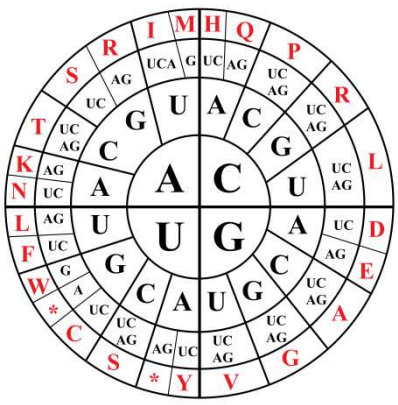

Fig. 3. Standard Codon Conversion Table

All the 64 codon combinations can be translated into a set of 20 essential amino acids, as seen in the standard codon table shown in Fig. 3. The 20 essential amino acids along with their one letter codes are - alanine (A), arginine (R), asparagine (N), aspartic acid (D), cysteine (C), glutamic acid (E), glutamine (Q), glycine (G), histidine (H), isoleucine (I), leucine (L), lysine (K), methionine (M), phenylalanine (F), proline (P), serine (S), threonine (T), tryptophan (W), tyrosine (Y), and valine (V). In this paper, the mRNA sequences for the 3 coronaviruses are translated into amino acids. Groups of amino acids form protein sequences, which are separated by the stop sign (*).

The amino acid composition of these translated proteins is presented graphically in Fig. 4. From Fig. 4, it is clear that all the 3 translated genomes have the highest composition of leucine (L) amino acid residues. The amino acid distribution for both bat-related coronaviruses is almost the same, with differences seen when compared with the SARS-CoV-2. The bat-derived coronaviruses bat-SL-CoVZC45 and bat-SL-CoVZXC21 have higher composition of alanine (A), glutamic acid (E), isoleucine (I), leucine (L), methionine (M) and valine (V) residues as compared to SARS-CoV-2. On the other hand, arginine (R), cysteine (C), phenylalanine (F), tryptophan (W) and tyrosine (Y) residues are in higher amounts in SARS-CoV-2. This difference in amino acid residues can be potentially used to distinguish the protein sequences obtained from these three coronaviruses.

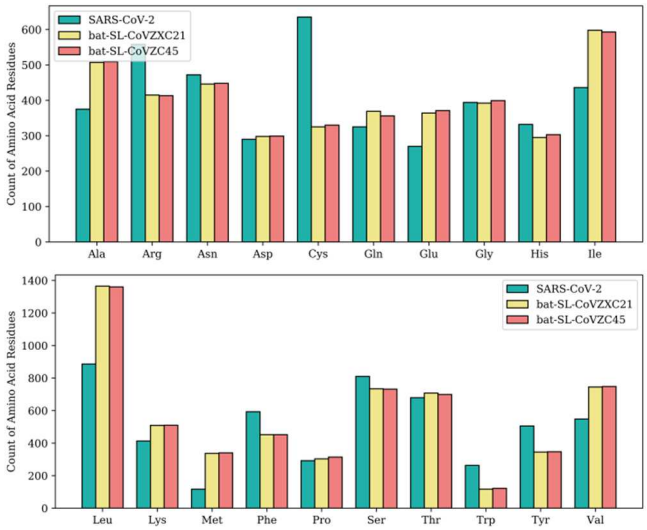


Fig. 4. Amino Acid Composition of the Three Coronaviruses

## VI. ANALYSIS OF STRUCTURAL PROTEINS

From the genbank files of the genomes, fasta sequences of the S, E, M and N proteins of the three coronaviruses were isolated. Using the ProtParam module in Biopython, 4 properties were calculated for these sequences:

### *A. Count of Amino Acid Residues:*

It is defined as the number of amino acid residues present in a protein sequence.

### *B. Aromaticity:*

It is defined as the sum of relative frequency of occurrence of the amino acids which are aromatic - Phenylalanine (F), Tyrosine (Y) and Tryptophan (W). This

method of calculation of aromaticity was first presented in [24].

### C. *Instability Index:*

It gives information on whether a protein will be stable in a test tube or not. A value greater than 40 indicates instability. The method of calculating the instability index was presented in [25].

### D. *Isoelectric Point:*

It is defined as the pH of the solution at which a protein carries no net electric charge and is calculated with the help of Bjellqvist methods given in [26-27].

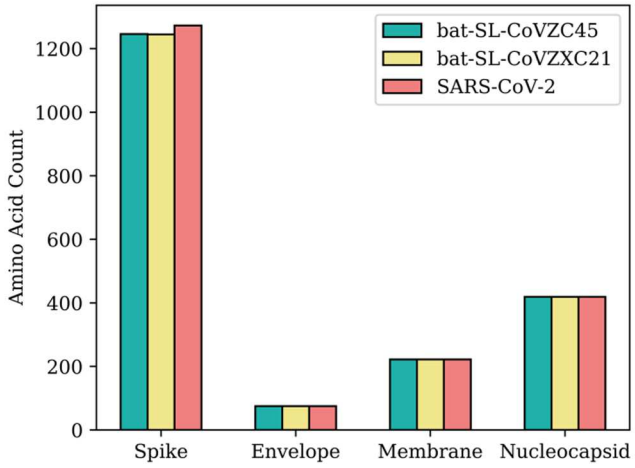


Fig. 5. Amino Acid Composition of the Structural Proteins

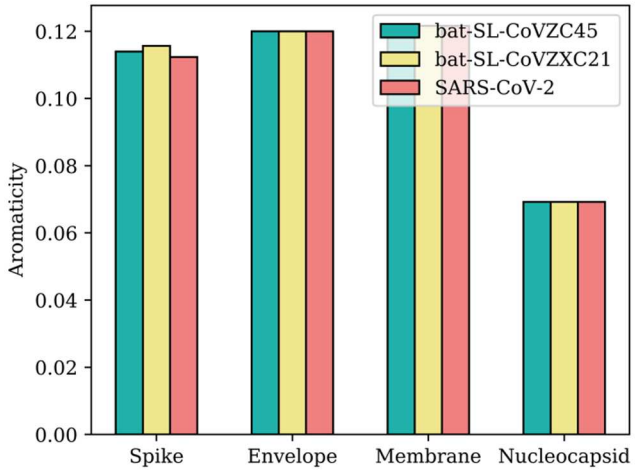


Fig. 6. Aromaticity Values of the Structural Proteins

The values of amino acid count, aromaticity, instability index and isoelectric point of the four structural proteins in the three coronaviruses are presented in Table IV, V, VI and VII respectively. These are also represented graphically for the sake of making conclusions easily in Figs. 5, 6, 7 and 8 respectively.

From Fig. 5, it can be seen that for all three coronaviruses, in terms of size, spike protein is the largest (has the largest amino acid count), followed by nucleocapsid protein, membrane protein and finally, envelope protein. They all have comparable (or in some cases, the same) length across the three coronavirus types.

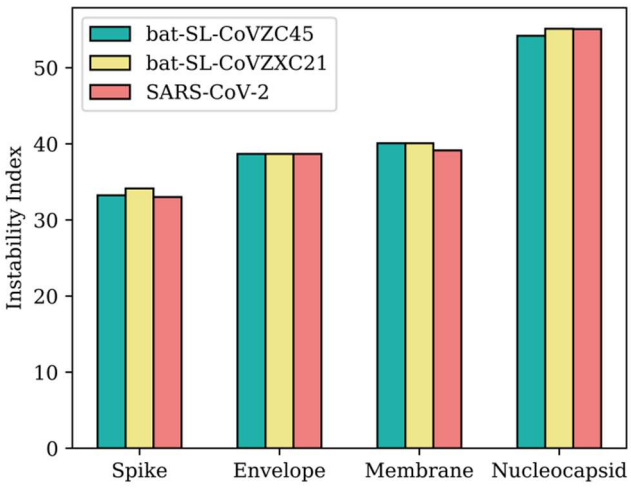


Fig. 7. Instability Index Values of the Structural Proteins

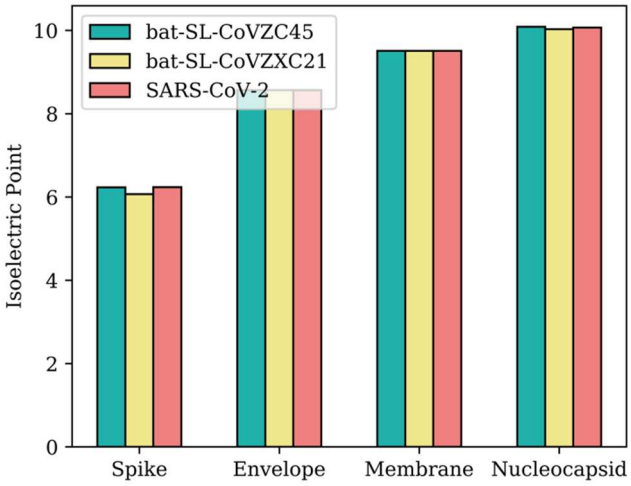


Fig. 8. Isoelectric Point Values of the Structural Proteins

TABLE IV. AMINO ACID COUNT IN THE STRUCTURAL PROTEINS

| Virus Strain | Total Amino Acid Count | | | |
|---|---|---|---|---|
| | ***S*** | ***E*** | ***M*** | ***N*** |
| SARS-CoV-2 | 1273 | 75 | 222 | 419 |
| bat-SL-CoVZXC21 | 1245 | 75 | 222 | 419 |
| bat-SL-CoVZC45 | 1246 | 75 | 222 | 419 |

TABLE V. AROMATICITY VALUES IN THE STRUCTURAL PROTEINS

| Virus Strain | Aromaticity | | | |
|---|---|---|---|---|
| | ***S*** | ***E*** | ***M*** | ***N*** |
| SARS-CoV-2 | 0.112 | 0.12 | 0.121 | 0.069 |
| bat-SL-CoVZXC21 | 0.115 | 0.12 | 0.121 | 0.069 |
| bat-SL-CoVZC45 | 0.114 | 0.12 | 0.121 | 0.069 |

From Fig. 6, it can be seen that aromaticity values are highest for membrane proteins, followed by envelope proteins, spike proteins and nucleocapsid proteins. The value of aromaticity is a good indicator of ringness of molecular compounds. Highly aromatic compounds (unsaturated, cyclic structures) show extra stability. This information can be of use by researchers to assess which out of the 4 proteins can be destabilized easily. These values are

also comparable across the 3 coronavirus types with minute variations in spike proteins.

TABLE VI. INSTABILITY INDICES OF THE STRUCTURAL PROTEINS

| Virus Strain | Instability Index | | | |
|---|---|---|---|---|
| | ***S*** | ***E*** | ***M*** | ***N*** |
| SARS-CoV-2 | 33.009 | 38.676 | 39.136 | 55.083 |
| bat-SL-CoVZXC21 | 34.141 | 38.676 | 40.074 | 55.116 |
| bat-SL-CoVZC45 | 33.236 | 38.676 | 40.074 | 54.196 |

TABLE VII. ISOELECTRIC POINTS OF THE STRUCTURAL PROTEINS

| Virus Strain | Isoelectric Point | | | |
|---|---|---|---|---|
| | ***S*** | ***E*** | ***M*** | ***N*** |
| SARS-CoV-2 | 6.236 | 8.568 | 9.509 | 10.069 |
| bat-SL-CoVZXC21 | 6.069 | 8.568 | 9.509 | 10.029 |
| bat-SL-CoVZC45 | 6.231 | 8.568 | 9.509 | 10.087 |

From Fig. 7, it is clear that the values of instability index are highest for N proteins, followed by M, E and S proteins. With values less than (or almost equal to) 40, the S, E and M proteins are stable. With values greater than 40, N proteins are unstable. Instability index values for the structural proteins across the 3 coronavirus types are comparable. Also from Fig. 8, it can be seen that the values of isoelectric points are highest for N proteins, followed by M, E and S proteins. Since these values give an indication of the pH ranges in which the protein is functional, it can be used by pharmaceutical companies to design drugs. Their values are also comparable for structural proteins across the 3 types.

A pairwise global sequence alignment between the 4 structural proteins was also done using EMBOSS Stretcher. The results of this alignment have been shown in Table VIII.

TABLE VIII. PID VALUES FOR STRUCTURAL PROTEINS

| Protein | Pair Aligned | PID Value |
|---|---|---|
| S | SARS-CoV-2/ bat-SL-CoVZXC21 | 80.0% |
| S | SARS-CoV-2/ bat-SL-CoVZC45 | 80.7% |
| M | SARS-CoV-2/ bat-SL-CoVZXC21 | 98.6% |
| M | SARS-CoV-2/ bat-SL-CoVZC45 | 98.6% |
| E | SARS-CoV-2/ bat-SL-CoVZXC21 | 100.0% |
| E | SARS-CoV-2/ bat-SL-CoVZC45 | 100.0% |
| N | SARS-CoV-2/ bat-SL-CoVZXC21 | 94.3% |
| N | SARS-CoV-2/ bat-SL-CoVZC45 | 94.3% |

The spike proteins in SARS-CoV-2 and bat-SL-CoVZC45 have a sequence identity of 80.7% and the spike proteins in bat-SL-CoVZXC21 and SARS-CoV-2 have a sequence identity of 80.0%. With sequence identity of 100%, it is seen that the envelope proteins in SARS-CoV-2, bat-SL-CoVZC45 and bat-SL-CoVZXC21 are completely identical. The membrane proteins show a sequence identity of 98.6% between bat-SL-CoVZXC21 & SARS-CoV-2 and bat-SL-CoVZC45 & SARS-CoV-2. Similarly, the nucleocapsid proteins show a sequence identity of 94.3% between bat-SL-CoVZXC21 & SARS-CoV-2 and bat-SL-CoVZC45 & SARS-CoV-2. The reason for low sequence identity in S proteins is attributed to the fact that the S protein in SARS-CoV-2 has modified due to homologous recombination of SARS-CoV and another Beta-CoV [28].

## VII. DISCUSSION AND CONCLUSION

In this paper, a data-based study was conducted to compare genome sequences of the novel coronavirus strain, SARS-CoV-2 and two bat-derived SARS-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, which have sequence identity very similar to the SARS-CoV-2. Additionally, their amino acid distributions were compared after translation. Also presented in this paper was a small analysis on the properties of the 4 structural proteins present in coronaviruses - the spike protein, membrane protein, envelope protein and nucleocapsid protein.

Results of global pairwise sequence alignment via the optimized Needleman-Wunsch Algorithm indicate that with a sequence identity of 87.8% between SARS-CoV-2 & bat-SL-CoVZC45 and 87.5% between SARS-CoV-2 & bat-SL-CoVZXC21, the sequences are similar in composition. This indicates that the novel coronavirus disease could be possibly linked to the disease in bats [10]. With a higher GC content, the bat-derived coronaviruses show somewhat higher stability than the SARS-CoV-2 on account of greater number of hydrogen bonds between G and C molecules. The dinucleotide composition between the 3 coronaviruses is almost similar, which is another proof to the fact that these genome sequences show good sequence identity.

Upon translation, the distributions of amino acid count revealed that bat-SL-CoVZC45 and bat-SL-CoVZXC21 have higher composition of alanine (A), glutamic acid (E), isoleucine (I), leucine (L), methionine (M) and valine (V) residues as compared to SARS-CoV-2. On the other hand, arginine (R), cysteine (C), phenylalanine (F), tryptophan (W) and tyrosine (Y) residues are in higher amounts in SARS-CoV-2.

An analysis on the properties of structural proteins such as amino acid count, aromaticity, instability index and isoelectric point revealed that there are little differences across the three coronavirus types. The S protein is the largest, followed by N, M and E proteins. If we talk about stability of molecular compounds imparted due to ringness and unsaturation, M proteins are the most stable, followed by E, S and N proteins. There is a general trend that is seen with instability index and isoelectric point values, with values highest for N proteins, followed by M, E and S proteins. The S proteins in SARS-CoV-2 and the bat-derived coronaviruses have 100% sequence identity, followed by M proteins with 98.6% and N proteins with 94.3%. The S proteins in SARS-CoV-2 and bat-SL-CoVZC45 have a sequence identity of 80.7% and the S proteins in bat-SL-CoVZXC21 and SARS-CoV-2 have a sequence identity of 80.0%. The author hopes that this

research will help future researchers and scientists in getting a better understanding of the virus strain and its differences from other bat-derived SARS-like coronaviruses.

## REFERENCES


[1] N. Zhong et al., "Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003," The Lancet, vol. 362, no. 9393, pp. 1353–1358, Oct. 2003.

[2] N. Wang et al., "Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4," Cell Research, vol. 23, no. 8, pp. 986–993, Jul. 2013.

[3] J. Zheng, "SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat," International Journal of Biological Sciences, vol. 16, no. 10, pp. 1678–1685, Mar. 2020.

[4] World Health Organisation (WHO). Accessed on: 8 November 2020. [Online]. Available: https://covid19.who.int

[5] S. Su et al., "Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses," Trends in Microbiology, vol. 24, no. 6, pp. 490–502, Jun. 2016.

[6] M. M. C. Lai and D. Cavanagh, "The Molecular Biology of Coronaviruses," in Advances in Virus Research, Elsevier, 1997.

[7] T. Hirano and M. Murakami, "COVID-19: A New Virus, but a Familiar Receptor and Cytokine Release Syndrome," Immunity, Apr. 2020.

[8] F. Wu et al., "A new coronavirus associated with human respiratory disease in China," Nature, vol. 579, no. 7798, pp. 265–269, Feb. 2020.

[9] M. Fahmi, Y. Kubota, and M. Ito, "Nonstructural proteins NS7b and NS8 are likely to be phylogenetically associated with evolution of 2019-nCoV," Infection, Genetics and Evolution, vol. 81, p. 104272, Jul. 2020.

[10] R. Lu et al., "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," The Lancet, vol. 395, no. 10224, pp. 565–574, Feb. 2020.

[11] S. Jiang, L. Du, and Z. Shi, "An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies," Emerging Microbes & Infections, vol. 9, no. 1, pp. 275–277, Jan. 2020.

[12] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges," International Journal of Antimicrobial Agents, vol. 55, no. 3, p. 105924, Mar. 2020.

[13] J. Zhang et al., "Insights into the cross-species evolution of 2019 novel coronavirus," Journal of Infection, vol. 80, no. 6, pp. 671–693, Jun. 2020.

[14] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," Journal of Advanced Research, vol. 24, pp. 91–98, Jul. 2020.

[15] R. N. Kirchdoerfer et al., "Pre-fusion structure of a human coronavirus spike protein," Nature, vol. 531, no. 7592, pp. 118–121, Mar. 2016.

[16] M. K. Gupta, S. Vemula, R. Donde, G. Gouda, L. Behera, and R. Vadde, "In-silico approaches to detect inhibitors of the human severe acute respiratory syndrome coronavirus envelope protein ion channel," Journal of Biomolecular Structure and Dynamics, pp. 1–11, Apr. 2020.

[17] S. Boopathi, A. B. Poma, and P. Kolandaivel, "Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment," Journal of Biomolecular Structure and Dynamics, pp. 1–10, Apr. 2020.

[18] National Center for Biotechnology Information. Accessed on: 15 May 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov

[19] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology, vol. 48, no. 3, pp. 443–453, Mar. 1970.

[20] E. W. Myers and W. Miller, "Optimal alignments in linear space," Bioinformatics, vol. 4, no. 1, pp. 11–17, 1988.

[21] F. Madeira et al., "The EMBL-EBI search and sequence analysis tools APIs in 2019," Nucleic Acids Research, vol. 47, no. W1, pp. W636–W641, Apr. 2019.

[22] A. C. . May, "Percent Sequence Identity," Structure, vol. 12, no. 5, pp. 737–738, May 2004.

[23] P. J. A. Cock et al., "Biopython: freely available Python tools for computational molecular biology and bioinformatics," Bioinformatics, vol. 25, no. 11, pp. 1422–1423, Mar. 2009.

[24] J. R. Lobry and C. Gautier, "Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes," Nucleic Acids Research, vol. 22, no. 15, pp. 3174–3180, 1994.

[25] K. Guruprasad, B. V. B. Reddy, and M. W. Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence," "Protein Engineering, Design and Selection," vol. 4, no. 2, pp. 155–161, 1990.

[26] B. Bjellqvist et al., "The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences," Electrophoresis, vol. 14, no. 1, pp. 1023–1031, 1993.

[27] B. Bjellqvist, B. Basse, E. Olsen, and J. E. Celis, "Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions," Electrophoresis, vol. 15, no. 1, pp. 529–539, 1994.

[28] B. Li et al., "Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing," mSphere, vol. 5, no. 1, Jan. 2020.