doi:10.20944/preprints202005.0022.v1

Structural and genetic analysis of coronaviruses spike proteins suggest pangolin as a proximate intermediate host of SARS-CoV-2 (COVID-19)

Sohail Raza[ζ1], Muhammad Asif Rasheed*[ζ2], Wajeeha Zahir[2], Muhammad Tariq Navid[3], Rana Aamir Diwan[4], Muhammad Awais[2], Tahir Yaqub[1], Masood Rabbani[1], Muhammad Rashid[5]

[1]Department of Microbiology, University of Veterinary and Animal Sciences, Lahore 54000, Pakistan
[2]Department of Biosciences, COMSATS University Islamabad, Sahiwal Campus, 57000, Sahiwal, Pakistan
[3] Department of Biological Sciences, National University of Medical sciences, Rawalpindi 46000, Pakistan
[4]Department of Community Medicine, Sahiwal Medical College, 57000, Sahiwal, Pakistan.
[5]Institute of Virology, Dr. Panjwani Center for Molecular medicine and Drug research (PCMD), International Center for Chemical and Biological Sciences (ICCBS), University of Karachi. 75270, Pakistan

* Corresponding Author
ζ Equally contributing authors

Abstract

During December 2019, a novel coronavirus named SARS-CoV-2 has emerged in Wuhan, China. The human to human transmission of this virus has also been established. The virus has so far infected more than 2 million people and spread over 200 countries. The World Health Organization (WHO) has declared COVID-19 a global health emergency due to its spread well beyond China. It has been established that this virus originates from bats and uses an intermediate host for transfer to humans. The knowledge about the intermediate host is important to find the virus shuttle mechanism to stop future outbreaks. For this, the genetic and structural analysis of coronaviruses spike proteins was performed using a computer-assisted approach.To conduct the *In silico* analysis, 43 sequences of spike protein belong to different species were retrieved from the NCBI nucleotide database. Pairwise and multiple sequence alignments were performed to check the similarities and differences of the retrieved sequences. Moreover, to highlight relationships among different species, phylogenetics analysis was performed using the MEGA software tool. In the end, protein structure alignment (superimposition) was performed against the reference structure by UCSF Chimera software. The results highlighted that the maximum similarity of human protein was found against Bat and Pangolinsequences. Moreover, among Bat and Pangolin, the highest similarity was found against pangolin based on phylogenetics analysis. These results suggest that SARS-CoV-2 transfers from bats to humans through pangolins.

doi:10.20944/preprints202005.0022.v1

1. Introduction

The current novel coronavirus disease 19 (COVID-19) emerged in Wuhan, China and captured the globe within four months, thus declaring global health emergency. During the last couple of decades, the coronaviruses have emerged with sever outbreaks i.e. Severe Acute Respiratory Syndrome coronaviruses (SARS-CoV) in 2003 and Middle East Respiratory Syndrome coronaviruses (MERS-CoV) in 2012. The COVID-19 is a third type of coronavirus that is inducing high morbidity and mortality in humans.

Since its initial outbreak in Wuhan, China, COVID-19 has infected more than 2.0 million of individuals with more than 0.12 million deaths and spread over 205 countries so far. The causative organism of COVID-19 is a *betacororonavirus* that is named as SARS-CoV-2. It is a zoonotic virus that transmits from animals to humans. Generally, the coronaviruses are enveloped viruses having a spherical or pleomorphic shape and contain positive-sense RNA. The genome is ranged between 26 to 32 Kbps with 80-120 nm of diameter [1]. The viral genome contains four structural proteins vis Envelop protein (E), Spike protein (S), Membrane protein (M) and Nucleo-capsid protein (N) [2]. Each of the proteins has an important role in virus life as S-protein for host cell attachment, N-protein for nucleocapsid formation, and M & E-proteins in viral assembly [3-5].

One of the important challenges is to determine the origin of SARS-CoV-2, to understand its transmission from animals to humans. Now it has been established that SARS-CoV-2 has been originated from horseshoe bats [6], but based on the transmission of earlier Coronavirus the bats cannot able to transfer the virus directly to the humans. For example, SARS-CoV, is the closest relative of SARS-CoV-2 that caused Severe Acute Respiratory Syndrome (SARS) pandemic in 2003. It was also transmitted from bats through intermediate host “the masked palm

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2020 doi:10.20944/preprints202005.0022.v1

civet" and finally to humans. Likewise, MERS, the Coronavirus that caused Middle East Respiratory Syndrome caused by Coronavirus in 2012, transmitted from bats to humans through intermediate host dromedary camel [7]. The human-animal interface is playing a major role in cross-species viral transmission and providing an appropriate place for various gene recombination; thus introducing new variants. The COVID-19 is supposed to be the product of the recombination of various coronaviruses that existed for a long time in various hosts and generate its ability to adapt to human hosts. However in this long journey, what kind of intermediate hosts have been used by coronaviruses to reach humans? It is entirely important to find the answer to broaden our understandings of the emergence and potential transmission of COVID-19. The identification of the intermediate host of SARS-CoV-2 is, therefore, a mystery that many researchers hope to solve, as knowing the intermediate host is very helpful for the prevention of the further spread of the epidemic.

In this study, to identify the intermediate host, the genetic and structural analyses of Coronaviruses (including SARS-CoV-2) spike proteins were performed using bioinformatics approaches. The results of genetic and structural analysis are predicted the pangolin as a proximate intermediate host of SARS-CoV-2.

2. Material and Method

2.1. Retrieval of glycoprotein sequences

The sequences of the spike protein of coronaviruses were retrieved from the National Center for Biotechnology Information (NCBI) nucleotide database. These sequences belong to different species including humans. Altogether, 43 sequences of different species were retrieved. The list of the species and the accession numbers are mentioned in **table 01**.

2.2. Pairwise alignment

The retrieved sequences were aligned to check the similarities and differences of the retrieved sequences against the reference sequence. For pairwise alignment, the human coronavirus sequence was taken as the reference sequence. All the retrieved sequences were aligned against the reference coronavirus sequence using the BLAST algorithm (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

2.3. Multiple sequence alignment

The multiple sequence alignment of all the retrieved sequences was performed to check the similarities and differences among the sequences. Moreover, multiple sequence alignment is required to perform phylogenetics analysis as well. The multiple sequence alignment of all the sequences was performed using Clustal Omega webserver (https://www.ebi.ac.uk/Tools/msa/clustalo/).

2.4. Phylogenetics analysis

To highlight the relationship among different species, phylogenetic analysis was performed by the Molecular Evolutionary Genetics Analysis (MEGA) 6.06software tool. Moreover, to confirm the results, phylogenetics analysis was performed by different algorithms including parsimony analysis, maximum likelihood analysis and unweighted pair group method with arithmetic mean (UPGMA) analysis.

2.5. Protein structure prediction and refinement

The protein structures of all the retrieved sequences mentioned in **table 01**were predicted either by homology modeling or threading algorithms. First of all, the templates for the sequences were searched in Protein Databank (PDB) database using the BLAST algorithm. The structures of all proteins were predicted by homology modeling where a good template was found in PDB. Moreover, the structures of the remaining proteins were predicted by threading where good templates of the proteins were not found.Furthermore; it was ensured that the quality of the structures is good. The qualities of the predicted protein structures were enhanced by Modrefiner (https://zhanglab.ccmb.med.umich.edu/ModRefiner/) web server for those proteins where quality was not so good.

2.6. Proteins' structure superimposition

To perform protein structure alignment, the superimpositions of the structures were performed by UCSF Chimera 1.14 software. Just like pairwise alignment, the predicted structures were aligned to check the similarities and differences of the predicted structures against the reference structure. The human coronavirus glycoprotein structure was taken as reference structure and all the predicted structures were compared against the reference structure.

3. Results

3.1. Pairwise Alignment

The pairwise alignment of all the retrieved sequences was performed against the reference human coronavirus sequence. The results of the pairwise sequence alignment are mentioned in **table 02.** According to the results, the sequence similarity against the reference sequence was found maximum in Bat1 specie while the least similarity was found in Bvirus4. The results were arranged based on the alignment score. Hence, Bat1 secured an 1843 score of alignment with 97% query coverage, 72.76% identity, and 0 E value. These figures highlight that the result is significant with minimum noise in the results. Moreover, Pangolin got the second-highest similarity score to compare to the reference human coronavirus sequence. Pangolin secured a 1620 score of alignment with 88% query coverage, 88.62% identity, and 0 E value. Furthermore, Bvirus4 got the least similarity against the reference sequence. Bvirus4 secured 30.4 scores of alignment with just 1% query coverage, 83.33% identity and 5.4 E value which highlights that the similarity of the sequence is very less compared to the reference sequence.

3.2. Multiple sequence alignment

The results of multiple sequence alignment are shown in **figure 1.** According to the results, the maximum length of the sequence was related to Canine3 (1481 amino acids) specie while the minimum length was related to Poultry 1 (224 amino acids) specie. Moreover, stars in the alignment highlight the conserved residues among all the compared species based on multiples sequence alignment.Although no conserved residues were found in multiple sequence alignment results when we compared all 43 sequences, some amino acid residues were conserved when we limit alignment to some species. For example, if we limit the species to Rat, Camel, and Bovine, then the sequences are conserved at many points. This highlights the similarity of coronavirus sequences among these three species. Hence, there is a possibility that Rat, Camel, and Bovine directly infected themselves during the transmission of coronavirus.

3.3. Phylogenetics analysis

The phylogenetic analysis was performed by MEGA 6.06 software. To confirm the results, the analysis was performed by three different algorithms including Parsimony, Maximum Likelihood, and UPGMA.The clusters in the tree were formed according to species. For example, bovines were present in one cluster, camels were present in another cluster. Moreover, rats, bats and canine were present in their respective cluster as shown in **figure 2**. Interestingly, all results clustered human, bat, and pangolin in one cluster. This cluster is of great significance as we are focusing on human coronavirus origin. Moreover, in this cluster, the human had a closer relationship with pangolin compare to Bat1. This highlights the possibility that the origin of human coronavirus is from Bat to Pangolin to Human.

3.4. Proteins' structures analysis

The predicted structures of all the proteins were compared against the reference structure of human coronavirus glycoprotein. The structure comparison is shown in **figure3 and figures 4**. According to the comparison, the maximum similarity among the structures was found in Pangolin species while Bats also got significant similarities against the Human glycoprotein structure. Moreover, if we compare the resemblance of structures among different species, then Pangolin and Bats got maximum similarities in the structure to compare to humans than the rest of the species. This result strengthens the results of pairwise alignment and phylogenetics analysis where Pangolin and Bats showed the highest similarities against humans.

4. Discussion

An earlier study claimed that snakes were likely to be the intermediate hosts of the SARS-CoV-2. The researchers compared the codon usage in the SARS-CoV-2 virus against that of the cells in eight animals at the Wuhan Huanan Seafood Wholesale Market. That study found that the snakes share the most similar codon usage pattern to SARS-CoV-2, thereby declaring that snakes were the most likely intermediate hosts [8]. A follow-up study compared the codon usages of three coronaviruses (SARS-CoV-2, SARS-CoV, and MERS-CoV) to those of more than 10,000 different kinds of animals, suggesting that the early claim of snake-borne transmission of SARS-CoV-2 is likely to be incorrect [9].

The spike protein of SARS-CoV-2 is the most important concerning viral infection. The spike protein is the outermost protein of the virus that involves the attachment of the virus to the cell receptors called Angiotensin Converting Enzyme 2 (ACE2). The ACE2 is the transmembrane receptor on the mammalian cells that utilize by the SARS-CoV-2 for infection. Therefore, the spike protein involves in the infectivity and host specificity of the coronaviruses, and the good target to find the possible origin of the SARS-CoV-2 [10]. In this study, 42 sequences of Coronaviruses spike protein from different host species were compared with the SARS-CoV-2 spike protein to find the possible intermediate host. The bats are still the probable host of origin for SARA-CoV-2 [11]. The multiple sequence alignment results of all spike protein against SARA-CoV-2 spike protein exhibited that pangolin Coronavirus spike protein has a maximum similarity of 88.62% with SARS-CoV-2 as compare to bat Coronavirus, which exhibited similarity of 72.76%. This sequence similarity of SARS-CoV-2 with pangolin Coronavirus is very high as the spike protein is the main protein that binds the cell receptor thus determines the host specificity. Moreover, there are only 5 amino acid sequences that are different on Receptor Binding Domain (RBD) of pangolin spike protein as compare to bat coronavirus, which has 19 different amino acids on Receptor Binding Domain (RBD) of the spike protein. Recently, a published study shows similar results and predicts pangolin as the intermediate host [9]. The phylogenetic analysis of spike protein using three different algorithms confirms the above findings that humans, bat, and pangolin coronaviruses were found in the same cluster, however; spike protein of pangolin virus has a close relation with human Coronavirus than bat coronavirus. These results suggest the possible role of pangolin as an intermediate host of transferring SARS-CoV-2 from bats to humans as suggested in **figure 5**. Similar results were reported in a recently published study [12]. To further examine our findings, we performed the structural analysis by comparing the spike protein structure of human Coronavirus with other coronaviruses. As previously predicted, the human Coronavirus spike protein is more closely related to the pangolin Coronavirus than bat or other coronaviruses. Taking all together, the MSA, phylogenetic analysis and structural analysis predict pangolin as an intermediate host. As the virus originates from the live food market Wuhan, where wild animals including bats and the pangolin were kept together that provide the best environment of Coronavirus transfer between hosts.

5. Conclusion

Amid the COVID-19 outbreak, the detailed understanding of how the SARA-CoV-2 transfers to humans will be helpful in the prevention of future outbreaks. The SARS-CoV-2 transfers from bats to humans through an intermediate host. Using the genetic and structural analysis of spike proteins from different coronaviruses, we predict that pangolins served as an intermediate host to transfer the novel virus from bats to humans.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

No acknowledgments.

References

1. T. S. Fung and D. X. Liu, "Human Coronavirus: Host-Pathogen Interaction," *Annu Rev Microbiol*, vol. 73, pp. 529-557, 2019.

2. D. Schoeman and B. C. Fielding, "Coronavirus envelope protein: current knowledge," *Virol J*, vol. 16, no. 1, pp. 69, 2019.

3. R. N. Kirchdoerfer, C. A. Cottrell, N. Wang, J. Pallesen, H. M. Yassine, H. L. Turner, K. S. Corbett, B. S. Graham, J. S. McLellan and A. B. Ward, "Pre-fusion structure of a human coronavirus spike protein," *Nature*, vol. 531, no. 7592, pp. 118-121, 2016.

4. R. McBride, M. van Zyl and B. C. Fielding, "The coronavirus nucleocapsid is a multifunctional protein," *Viruses*, vol. 6, no. 8, pp. 2991-3018, 2014.

5. P. S. Masters, "The molecular biology of coronaviruses," *Adv Virus Res*, vol. 66, pp. 193-292, 2006.

6. P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L. Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shen, X. Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G. F. Xiao and Z. L. Shi, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270-273, 2020.

7. C. B. Reusken, V. S. Raj, M. P. Koopmans and B. L. Haagmans, "Cross host transmission in the emergence of MERS coronavirus," *Curr Opin Virol*, vol. 16, pp. 55-62, 2016.

8. W. Ji, W. Wang, X. Zhao, J. Zai and X. Li, "Cross-species transmission of the newly identified coronavirus 2019-nCoV," *J Med Virol*, vol. 92, no. 4, pp. 433-440, 2020.

9. C. Zhang, W. Zheng, X. Huang, E. W. Bell, X. Zhou and Y. Zhang, "Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1," *J Proteome Res*, vol. 19, no. 4, pp. 1351-1360, 2020.

10. M. Letko, A. Marzi and V. Munster, "Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses," *Nat Microbiol*, vol. 5, no. 4, pp. 562-569, 2020.

11. T. Zhang, Q. Wu and Z. Zhang, "Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak," *Curr Biol*, 2020.

12. J. Zhang, W. Jia, J. Zhu, B. Li, J. Xing, M. Liao and W. Qi, "Insights into the cross-species evolution of 2019 novel coronavirus," *J Infect*, 2020.

Table 01: The sequences of spike protein retrieved from NCBI nucleotide database. Altogether, 43 sequences of different species were retrieved.

Sr No.	ACCESION NO	ORGANISM
1	JF792617	Rat coronavirus
2	JF792616	Rat coronavirus
3	NC_012936	Rat coronavirus
4	FJ938068	Rat coronavirus
5	NC_032730	Rat coronavirus
6	KF294380	Rat coronavirus
7	KT368891	Camel coronavirus
8	JF792615	Dromedary camel coronavirus
9	JF792614	Dromedary camel coronavirus
10	NC_012937	Dromedary camel coronavirus
11	KT368892	Bovine coronavirus
12	JF792613	Bovine coronavirus
13	JF792612	Bovine coronavirus
14	MH043953	Bovine coronavirus
15	MH043952	Bovine coronavirus
16	AF220295	Bovine coronavirus
17	AF353511	Porcine epidemic diarrhea virus
18	KP890336	Porcine epidemic diarrhea virus
19	MG546690	Porcine epidemic diarrhea virus
20	MG546687	Porcine epidemic diarrhea virus
21	MF807952	Porcine epidemic diarrhea virus
22	MF807951	Porcine epidemic diarrhea virus
23	MF782687	Porcine epidemic diarrhea virus
24	KF663561	Infectious bronchitis virus
25	KF663560	Infectious bronchitis virus
26	KF663559	Infectious bronchitis virus
27	KC008600	Infectious bronchitis virus
28	KX272465	Infectious bronchitis virus
29	MK878536	Infectious bronchitis virus
30	KP981644	Canine coronavirus
31	GQ477367	Canine coronavirus
32	AY307021	Canine coronavirus
33	AY307020	Canine coronavirus
34	KY938558	Bat coronavirus
35	NC_009988	Bat coronavirus

36	EF203067	Bat coronavirus
37	EF203066	Bat coronavirus
38	EF203065	Bat coronavirus
39	EF203064	Bat coronavirus
40	KX442565	Bat coronavirus
41	KX442564	Bat coronavirus
42	QHD43416	Human coronavirus
43	[9]	Pangolin coronavirus

Table 02: The pairwise alignment results of all the retrieved sequences. The alignment was performed against the reference human coronavirus sequence. The maximum similarity was found in Bat1 specie while the least similarity was found in Bvirus4. The data was arranged on the basis of alignment score.

Sr No.	Organism	Total Score	Query Cover	E value	Identity
1	bat1	1843	97%	0	72.76%
2	pangolin	1620	88%	0	88.62%
3	bovine3	579	73%	1.00E-151	37.68%
4	bovine4	578	73%	2.00E-151	37.68%
5	bovine2	578	73%	3.00E-151	37.68%
6	bovine5	577	73%	5.00E-151	37.68%
7	bat7	574	79%	0	35.94%
8	bovine1	573	73%	1.00E-148	37.55%
9	bovine6	569	73%	2.00E-147	37.15%
10	camel4	568	75%	2.00E-146	36.63%
11	camel3	568	91%	2.00E-146	36.50%
12	camel2	568	91%	2.00E-146	36.50%
13	camel1	567	91%	3.00E-146	36.50%
14	bat8	561	79%	5.00E-180	35.03%
15	rat5	545	71%	1.00E-140	36.69%
16	rat2	545	71%	1.00E-140	36.69%
17	rat4	541	71%	3.00E-139	36.30%
18	rat3	541	71%	3.00E-139	36.30%
19	rat1	541	71%	3.00E-139	36.30%
20	rat6	381	79%	3.00E-107	29.42%
21	Bvirus6	373	44%	2.00E-102	37.31%
22	bat6	367	63%	8.00E-101	31.14%

23	bat2	367	63%	8.00E-101	31.14%
24	virus7	367	66%	3.00E-92	30.33%
25	bat5	365	63%	2.00E-100	31.23%
26	bat4	365	63%	3.00E-100	31.23%
27	bat3	365	63%	3.00E-100	31.23%
28	virus1	365	65%	9.00E-91	30.44%
29	Bvirus5	357	39%	1.00E-106	38.36%
30	virus5	349	60%	7.00E-93	30.33%
31	virus3	347	58%	8.00E-93	30.33%
32	virus4	347	60%	2.00E-92	30.16%
33	canine4	345	56%	5.00E-101	31.61%
34	canine3	342	56%	7.00E-100	30.80%
35	canine1	340	56%	3.00E-99	32.07%
36	virus6	340	60%	3.00E-90	29.63%
37	canine2	339	56%	7.00E-99	31.82%
38	virus2	337	60%	1.00E-89	30.35%
39	Bvirus1	31.6	1%	3.7	66.67%
40	Bvirus2	31.6	1%	3.6	66.67%
41	Bvirus3	30.4	1%	5.4	83.33%
42	Bvirus4	30.4	1%	5.4	83.33%

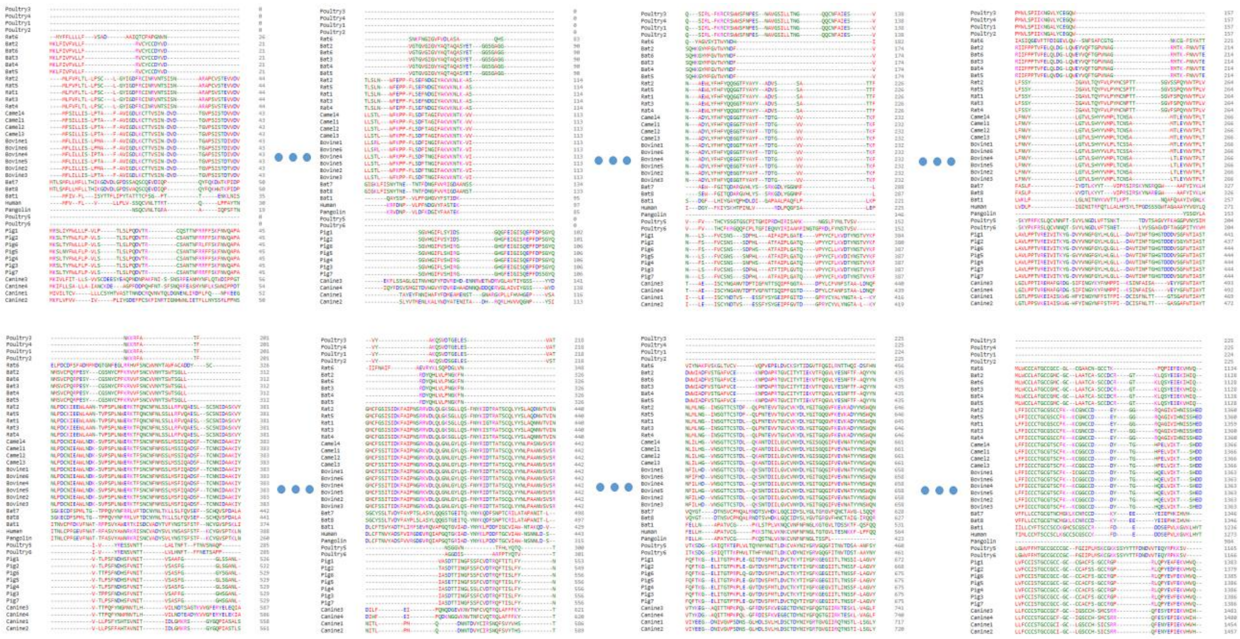

Figure 1: The results of multiple sequence alignment**.** The maximum length of the sequence was found in Canine3 with 1481 amino acids while specie while minimum length was found in Bvirus1 with 224 amino acids specie.

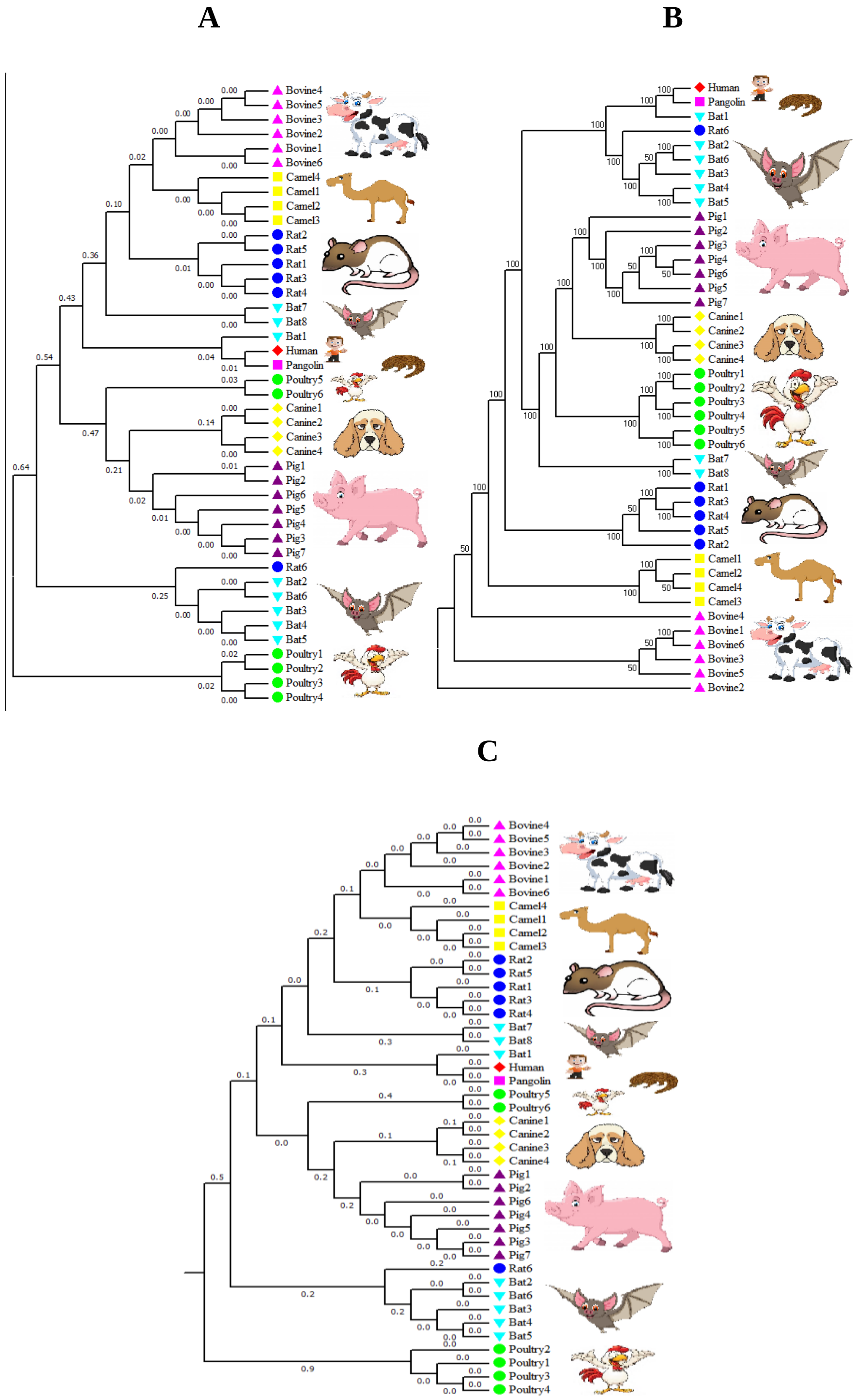

Figure 2: The phylogenetics analysis of all species included in the current study. The results were confirmed by three different algorithms including Maximum Likelihood (A), Maximum Parsimony (B) and UPGMA (C). All results highlighted that human corona virus has maximum relation with pangolin. The analysis was performed by MEGA 6.06 software.

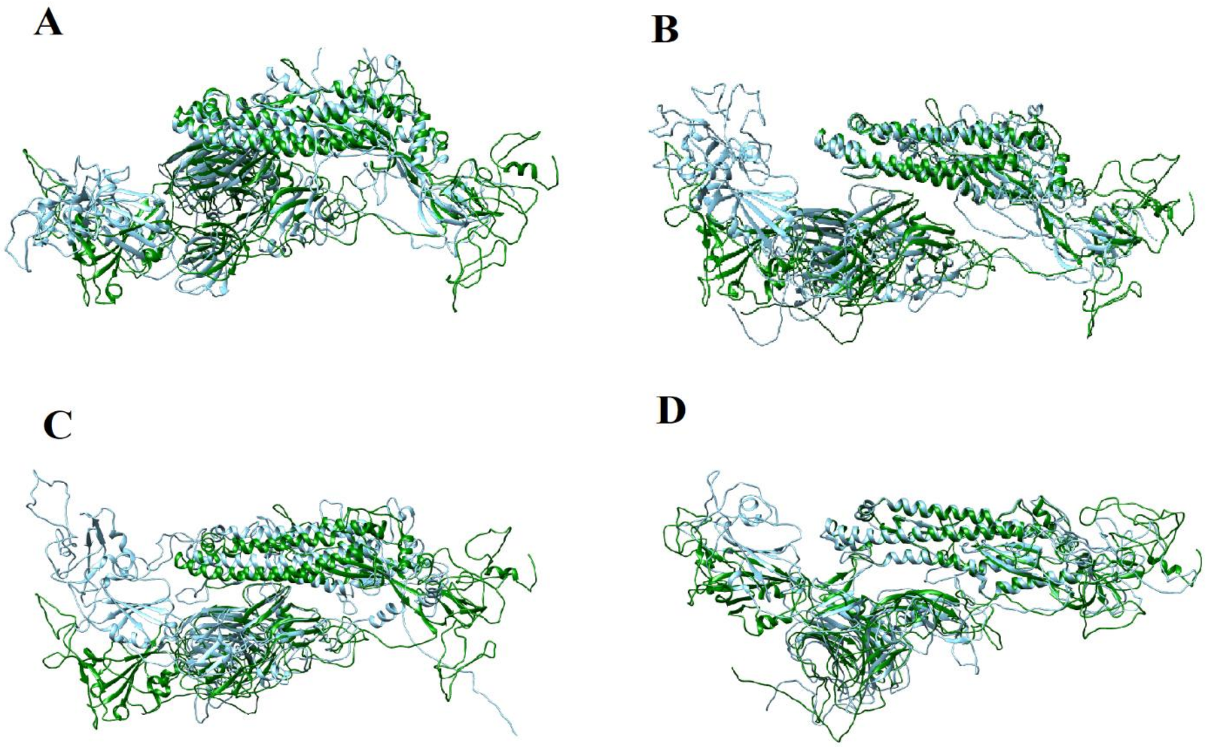

Figure 3: Comparative structures analysis for spike protein structures of all the species in present study against human SARS-CoV-2 spike protein (structure shown in green). The predicted structures of all the proteins were compared against reference structure of human corona virus glycoprotein by superimposition. A = Bovine 1. B = Camel 1. C = Rat1. D = Bat1. Amongst these species, the maximum similarity among the structures was found in Bat1.

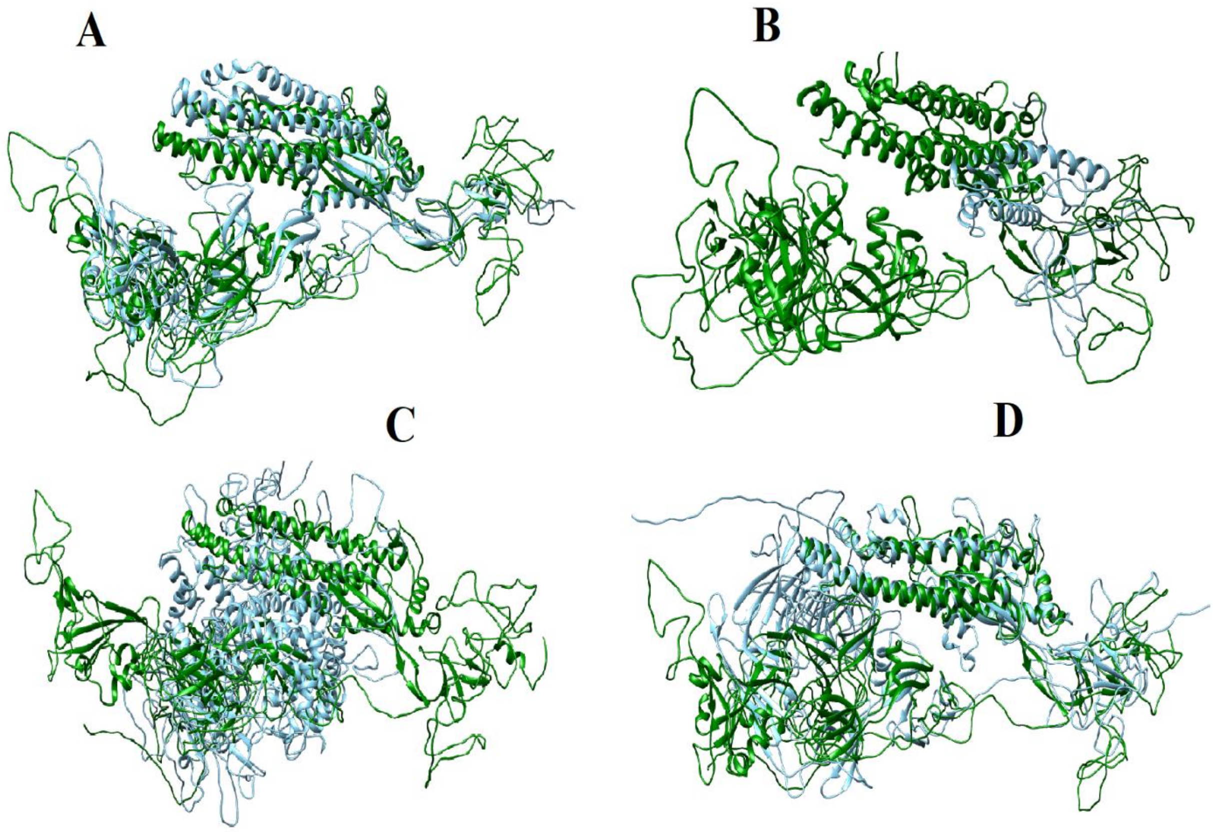

Figure 4: Comparative structures analysis for spike protein structures of all the species in present study against human SARS-CoV-2 spike protein structure (structure shown in green). The predicted structures of all the proteins were compared against reference structure of human corona virus glycoprotein by superimposition. A = Pangolin. B = Poultry 1. C = Canine 1. D = Pig 1. Amongst these species, the maximum similarity among the structures was found in Pangolin.

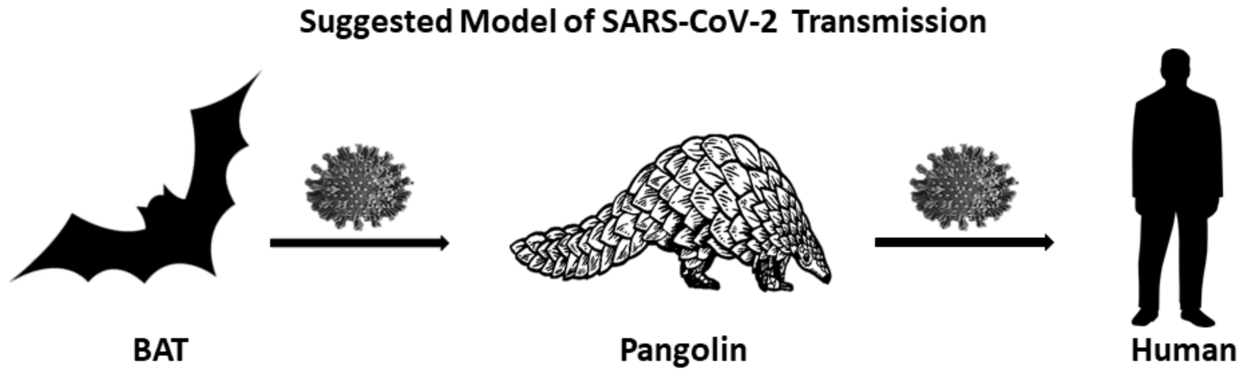

Figure 5: The results of present study suggest the possible role of pangolin as an intermediate host of transferring SARS-CoV-2 from bats to humans.