[Frontiers in Bioscience, Landmark, 25, 1894-1900, June 1, 2020]

Comparative analysis of *Coronaviridae* nucleocapsid and surface glycoprotein sequences

Babu V. Bassa[1], Olen R. Brown[2]

[1]Southern University and A and M College, Baton Rouge, LA 70813 [2]University of Missouri Dalton Cardiovascular Research Center, Columbia, MO 65211

TABLE OF CONTENTS

1. ABSTRACT

We analyzed the nucleocapsid and surface proteins from several *Coronaviridae* viruses using an alignment-free computer program. Three isolates of novel, human coronavirus (SARS0CoV-2) (2019) that are responsible for the current pandemic and older SARS strains of human and animal coronaviruses were examined. The nucleocapsid and glycoprotein sequences are identical for the three novel 2019 human isolates and they are closely related to these sequences in six bat and human SARS coronaviruses. This strongly supports the bat origin of the pandemic, novel coronavirus. One surface glycoprotein fragment of 111 amino acids is the largest, conserved, common permutation in the examined bat SARS-like and human SARS viruses, including the Covid-19 virus. BLAST analysis confirmed that this fragment is conserved only in the human and bat SARS strains. This fragment likely is involved in infectivity and is of interest for vaccine development. Surface glycoprotein and nucleocapsid protein sequence homologies of 58.9% and 82.5%, respectively, between the novel SARS0CoV-2 strains and the human SARS (2018) virus suggest that existing anti-SARS vaccines may provide some protection against the novel coronavirus.

2. INTRODUCTION

Viruses belonging to the family *Coronaviridae* are known to cause severe and acute lung inflammation in humans and other animals (1). Comparative analysis of coronavirus proteins is useful for understanding the relationships of these viruses with respect to their origins, for developing more specific diagnostic tests, and to design vaccines against the novel coronavirus that is causing widespread morbidity and mortality across the world. We used an alignment-free software program (Compare) developed by one of us (Babu V. Bassa) for comparing surface glycoprotein and nucleocapsid proteins of the coronaviruses. Non-alignment programs are considered to be superior to the alignment programs because of known uncertainties associated with the alignment of sequences (2). Our program extracts common amino acid sequences (permutations) that are five residues or larger from any given pair of proteins. This procedure identifies conserved fragments and provides information on the physical similarities among the primary structures of biological sequences.

3. METHODS

The analysis was done using the unique software tool “Compare”. The algorithm was implemented in Microsoft’s Visual Basic language for the Windows Operating System. An outline of the algorithm is presented in Figure 1. The source code

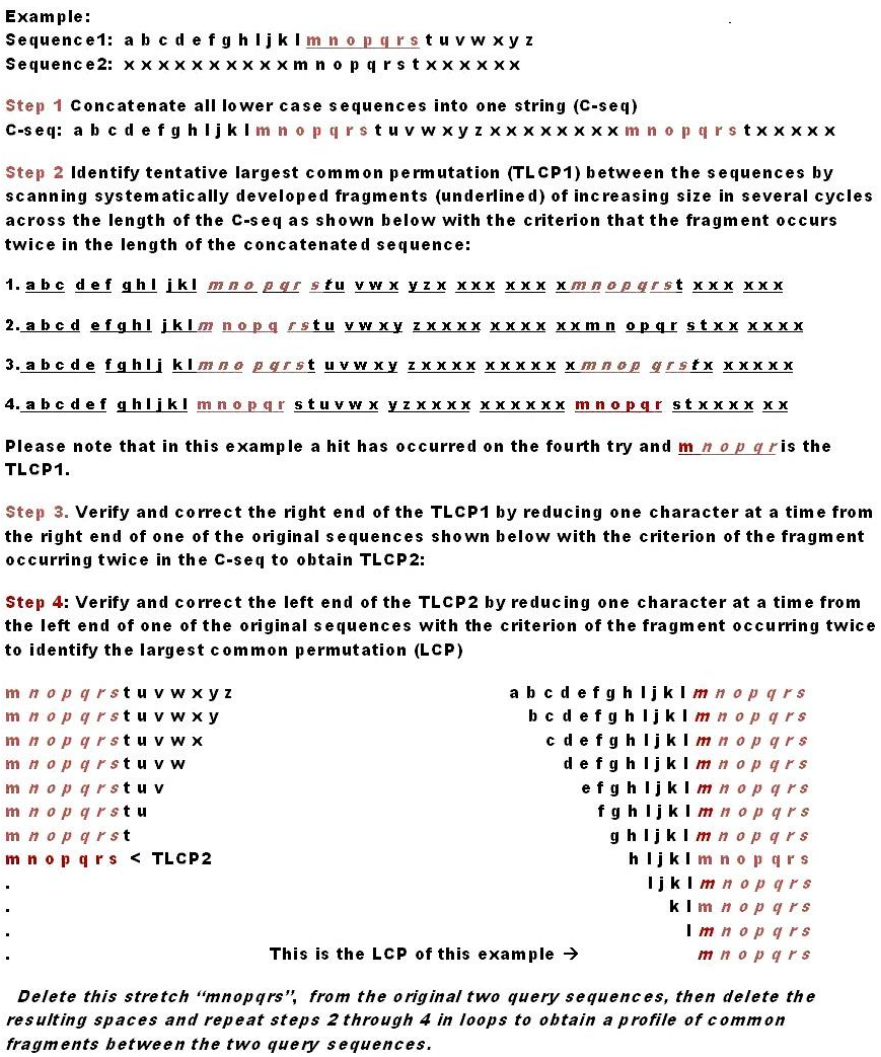

Figure 1. This example shows how the common permutation of the two sequences ("mnopqrs") is identified.

and the raw data will be made available to the Journal for distribution.

In keeping with the current terminology, the coronavirus strain of the current outbreak is referred to as "novel coronavirus" and the strains prior to the 2019 outbreak that caused severe acute respiratory syndrome are referred to as "SARS strains" throughout the manuscript. Similarly bat SARS viruses are referred to as "bat SARS-like strains".

The coronavirus sequences used in this analysis were obtained from the GeneBank. The gene bank accession numbers for all comparisons are given in Table 1. Additionally, a large number of animal coronaviruses were screened and were found to be very distant in terms of sequence homologies (described later in this section) to the novel coronavirus. The severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01, GeneBank- Accession number: MN988668 (11-FEB-2020), was obtained as the complete genome (3). The severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, GeneBank-Accession number: NC-045512.2 (13-MAR-2020), was obtained as the

© 1996-2020

Table 1. Nucleocapsid and surface glycoprotein sequence relationships among various strains of coronaviruses

Coronavirus	SARS0CoV-2 AC: MT072688				
	Nucleocapsid protein		Surface glycoprotein		Accession
	%Homology	Largest Common Permutation	%Homology	Largest Common Permutation	
SARS0CoV-2(1) Feb 2020	100	419	100	1273	MN988668
SARS0CoV-2(2) Mar 2020	100	419	100	1273	NC_045512
Bat.RaTG13 Mar 2020	99	177	99	440	MN996532
Rhinophol. affins-2014	85.4	99	59.2	111	KF569996
SARS.BAT-July2017	83.7	66	57.9	111	JX993988
SARS.Human Aug 2018	82.5	43	58.9	111	NC_004718
SARS.Bat Dec 2017	82.5	43	59.4	111	KY417152
Kenyan Bats Feb 2020	73.3	44	50.9	97	KY352407
M.East.Res.Syndrome May 2016	16.6	8	5.4	14	KX034100
Avian Aug 2018	8.2	6	3.1	8	NC_001451
Human. Dec.2018	6.5	12	4.8	7	NC_003045
Bat.Corona.HKUB Apr 2008	1	4	2.1	8	EU420139
Actual fragments are shown in Figures 1-3. Viruses are arranged in decreasing order of homology with the novel coronavirus					

complete genome (4). The SARS coronavirus, GeneBank, Accession number: NC_004718 (13-AUG-2018), was obtained as the complete genome (5).

Several-hundred whole, genome sequences from the family *Coronaviridae* are available in the GeneBank; however, many were found to be repeats of the same strains. In the initial phase of this study we screened multiple combinations (more than 100) of nucleocapsid proteins and surface glycoproteins from various coronaviruses. The three novel coronavirus isolates (SARS0CoV-2) were identical and they were similar to SARS and bat SARS-like viruses. Based on this initial screening, pairs of viral strains were chosen for final analyses. Nucleocapsid protein (NCP) and surface glycoprotein (SGP) were selected for comparative analyses because of their known importance in infection and in the immune response. To calculate sequence homologies, the character lengths of the common fragments equal to or larger than five amino acid residues were summed and the percentages were computed based on the total sequence lengths. The homology parameter so obtained is a relative index applicable only to this method of calculation.

4. RESULTS

Protein fragments from the three isolates of novel coronavirus (SARS0CoV-2) are identical and they have higher degrees of sequence homology with SARS and bat SARS-like strains reported prior to the current outbreak (Table 1). The latest reported bat strain, Bat.RaTG13-Mar-2020, (6) has 99% sequence homology with the novel coronavirus with respect to both SGP and NCP. The Kenyan bat coronavirus genome that was deposited in 2016 by the Centers for Disease Control laboratory, Atlanta Georgia (7), has significant homology with novel coronavirus with respect to both NCP and SGP (Table 1). There is a largest common permutation (111 residues long) that is conserved in the SGP of novel coronavirus, SARS strains, and bat SARS-like coronaviruses. It is part of the 440 fragment of Bat.RaTG13 (Table 1). Therefore, the 111-residue is present in the novel coronavirus, in at least one human SARS strain of coronavirus, and in at least one bat SARS-

© 1996-2020

A .

1.rsvasqsiiaytmslgaensvaysnnsiaiptnftisvtteilpvsmtktsvdctmyicgdstecsnlllqygsfctqlnr altgiaveqdkntqevfaqvkqiyktppikdfggfnfsqilpdpskpskrsfiedllfnkvtladagfikqygdclgdiaar dlicaqkfngltvlpplltdemiaqytsallagtitsgwtfgagaalqipfamqmayrfngigvtqnvlyenqklianqfns aigkiqdslsstasalgklqdvvnqnaqalntlvkqlssnfgaissvlndilsrldkveaevqidrlitgrlqslqtyvtqqlir aaeirasanlaatkmsecvlgqskrvdfcgkgyhlmsfpqsaphgvvflhvtyvpaqeknfttapaichdgkahfpre gvfvsngthwfvtqrnfyepqiittdntfvsg(440)685

2.krfdnpvlpfndgvyfasteksniirgwifgttldsktqsllivnnatnvvikvcefqfcndpflgvyyhknnkswmes efrvyssannctfeyvsqpflmdlegkqgnfknlrefvfknidgyfkiyskhtpinlvrdlp(141)77

3.gfsaleplvdlpiginitrfqtllalhrsyltpgdsssgwtagaaayyvgylqprtfllkynengtitdavdcaldplsetkc tlksftvekgiyqtsnfrvqpt(105)219

4.cdvvigivnntvydplqpeldsfkeeldkyfknhtspdvdlgdisginasvvniqkeidrlnevaknlneslidlqelg kyeqyikwpwyiwlgfiagliai(102)1126

5.apatvcgpkkstnlvknkcvnfnfngltgtgvltesnkkflpfqqfgrdiadttdavrdpqtleilditpcsfggvsvitp gtn(84)520

6.snqvavlyqdvnctevpvaihadqltptwrvystgsnvfqtragcligaehvnnsyecdipigagicasyqtqtns(7 6)605 7.mvtimlccmtsccsclkgccscgscckfdeddsepvlkgvklhyt(45)1229

8.gdevrqiapgqtgkiadynyklpddftgcviawns(35)404 9.mfvflvllplvssqcvnlttrtqlppaytns(31)1 10.sfstfkcygvsptklndlcftnvyadsfvi(30)373 11.fasvyawnrkrisncvadysvlyns(25)373 12.tqdlflpffsnvtwfhaihvsgtng(25)51 13.sivrfpnitnlcpfgevfnat(21)325 14. nlkpferdisteiyqags(18)460 15. trgvyypdkvfrssvlh(17)33 16.qpyrvvvlsfell(13)506 17.nylyrlfrk(9)450 18.pcng (4)479

B .

1.algklqdvvnqnaqalntlvkqlssnfgaissvlndilsrldkveaevqidrlitgrlqslqtyvtqqliraaeirasanlaat kmsecvlgqskrvdfcgkgyhlmsfpq(111)944

2.nntvydplqpeldsfkeeldkyfknhtspdvdlgdisginasvvniqkeidrlnevaknlneslidlqelgkyeqyikw pwy(82)1134 3.gwtfgagaalqipfamqmayrfngigvtqnvlyenqk(37)885

4.cscgscckfdeddsepvlkgvklhyt(26)1248 5.ardlicaqkfngltvlppllld(22)846

6.pqiittdntfvsgncdvvigi(21)1112 7.vrfpnitnlcpfgevfnat(19)327

8.syecdipigagicasy(16)659 9.krsfiedllfnkvtladagf(20)814 10.wlgfiagliaivmvti(16)1217

11.nlllqygsfctqlnral(17)751 12.fggfnfsqilpdp(13)797 13.cvnfnfngltgtgvlt(16)538

14.isncvadysvlyns(14)358 15. aphgvvflhvtyvp(14)1056 16.gyqpyrvvvlsfell(15)504

17.lccmtsccsclkg(13)1234 18.nfttapaich(10)1074 19.agcligaehv(10)647

20.kgiyqtsnfrv(11)310 21.fggvsvitpgtn(12)592 22.engtitdavdc(11)281 23. iadynyklpddf(12)418 24.evfaqvkq (8)780 25.vavlyqdvnct(11)608 26.myicgdstec(10)740

27.fstfkcygvs (10)374 28. pfqqfgrd(8)561 29.apatvcgpk(9)520 30.vrqiapgqtg(10) 407

31.aihadqltp(9)623 32.tqdlflpf(8)51 33.tklndlcf(8)385

34.aytmslga(8)694 35.lrefvfkn(8)189 36.nvyadsfv(8)394 37.fpregvfv(8)1089 38.pferdis (7)463 39 rgvyypd (7)34 40. nctfey (6)165 41.ktsvdc(6)733 42.ianqfn(6)923 43.iaiptnf (7)712 44.vrdlp (5)213 45.clgdi (5)840 46.svyaw(5)349 47.eildi(5)583 48.tqrnf (5)1105

49.ginit (5)232 50.aysnn (5)706 51.teksn(5)95 52.kqyg(4)835

Figure 2. For each numbered motif, the size is given in parenthesis and the location of the fragment in the molecule is indicated by the underlined residue number.

like strain of coronavirus (Table 1). Based on this observation we have subjected the 111-fragment to a BLAST search and found that the 111-fragment is preserved only in SARS viruses (data not presented). The 111 SGP motif is absent in avian, MERS, some human, and some bat strains of coronaviruses. The NCP in corona viruses is only 419 amino acids long. As shown in Figure 3 and Table 1 there are several polypeptide motifs originating from this protein that are common to novel coronavirus, SARS and bat SARS- like strains of coronavirus. A compilation of common polypeptide motifs is presented in Figure 1, Figure 2, and Figure 3. These polypeptide motifs will be useful as detection tools in studying the origins of novel coronavirus. They also will be helpful in designing vaccine candidates.

5. DISCUSSION

Unlike alignment-based sequence comparison programs, our software tool allows

© 1996-2020

A.

1.kqrrpqglpnntaswftaltqhgkedlkfprgqgvpintnsspddqigyyrratrrirggdgkmkdlsprwyfyylgtg peaglpygankdgiiwvategalntpkdhigtrnpannaaivlqlpqgttlpkgfyaegsrggsqassrsssrsrnssrn stpgssrgtsparmagng (177)38
2.ynvtqafgrrgpeqtqgnfgdqelirqgtdykhwpqiaqfapsasaffgmsrigmevtpsgtwltytgaiklddkdpn fkdqvillnkhidayktfpptepkkdkkkkadetqalpqrqkkqqtvtllpaadlddfskqlqqsmssadstqa (152)268 3.msdngpqnqrnapritfggpsdstgsnqngersgar(36)1 4.qtvtkksaaeaskkprqkrtatk (23)244 5.daalalllldrlnqleskmsgkgqqqq (27)216

B.

1. lirqgtdykhwpqiaqfapsasaffgmsrigmevtpsgtwlty(43) 291
2. sgkgqqqqgqtvtkksaaeaskkprqkrtatk(32) 235
3. vlqlpqgttlpkgfyaegsrggsqassrsssrsr(34)158
4. kqrrpqglpnntaswftaltqhgke(25) 38 5.villnkhidayktfpptepkkdkkkk(26)350
6. ynvtqafgrrgpeqtqgnfgdq (22)268 7.wvategalntpkdhigtrnp (20)132
8. lsprwyfyylgtgpea (16)104 9.alalllldrlnqlesk (16) 218
10.pddqigyyrratrr(14)80 11.srnstpgssrg (11)194 12.fprgqgvpintns (13)66
13. apritfggp (9)12 14. gaikldkdp (10) 335 15. tvtllpaad (9)391 16. sadstqa(7)413
17.lpqrqkkq (8)382 18.msdngpq(7)1 19.rggdgkmk(8)95 20.lpygank (7)121
21.sparma(6)206 22.ddfs (4)401

Figure 3. For each numbered motif, the size is given in parenthesis and the location of the fragment in the molecule is indicated by the underlined residue number.

A.

1.alntlvkqlssnfgaissvlndilsrldkveaevqidrlitgrlqslqtyvtqqliraaeirasanlaatkmsecvlgqskrvdfcgkg yhlmsfpq(97)958 2.tfgagaalqipfamqmayrfngigvtqnvlyenqk (35)887
3.myicgdstecsnlllqygsfctqlnral (28)740 4. ardlicaqkfngltvlpplltd (22)846
5.lvknkcvnfnfngltgtgvlt(21)533 6. aknlneslidlqelgkyeqy(20)1190 7.sfstfkcygvsptklndlcf(20)373
8.iadynyklpddftgcv(16)418 9.ikdfggfnfsqilpdp(16)794 10.ggvsvitpgtnts(13)593
11.eldkyfknhtspdvd(15)115112.gvgyqpyrvvvlsfell(17)502 13.tvekgiyqtsnfrv(14)307 14. aphgvvflhvtyvp(14)1056 15.Yecdipigagica(13) 660 16. apatvcgpkkst(12) 520
17.vrfpnitnlcpfg(13) 327 18. gdisginasvv(11)1167 19. cscgscckfded (12)1248 20. lpvsmtktsvdc(12)727 21. vavlyqdvnct(11)608 22. engtitdavdc(11)281 23. nfttapaich(10)1074
24.gvklhyt(7)1267 25.nntvydplqpel(12)1134 26.fpregvfv(8)1089 27.algklqdv(8)944
28.kqygdclg(8)835 29.rsfiedll (8) 815 30.cdvvigi(7)1126 31.pfqqfgrd(8)561 32.iaiptnf(7)712
33.giaveqd(7)769 34.imlccmt(7)1232 35.ittdntf(7)1115 36.kiqdsl(6)933 37.tqrnfy(6)1105
38.ngthwf(6)1098 39. vlyns(5)367 40. isvtte(6)720 41. irgwifg(7)101 42.vrqiap(6) 407 43.iyktp(5)788
44.cvady(5)361 45.kwpwy(5)1211 46.aytms(5)694 47.frssv(5)43 48.eildi(5)583 49.nkvtl(5)824
50.ianqfn(6)923 51.svyaw(5)349 52.erdis(5)465 53.wlgf(4)1217

B.

1.kwpwyiwlgfiagl(14)1211 2.enqklian(8)918 3.ardlicaq(8)846
4.iedllf(6)818 5.qidrli(6)992 6.naqal(5)955 7.lvllp(5)5
8.gwtag(5)257 9.iptnf(5)714 10.vlppl(5)860 11.gvtq(4)910

Figure 4. For each numbered motif, the size is given in parenthesis and the location of the fragment in the molecule is indicated by the underlined residue number.

comparison of sequences by identifying and making profiles of common permutations between given pairs of biological sequences. The picture captured is easily understood and interpreted and does not have some uncertainties associated with alignment-based sequence comparison programs. The size of the largest common permutation is an easily understood parameter of the relationships among the sequence

© 1996-2020

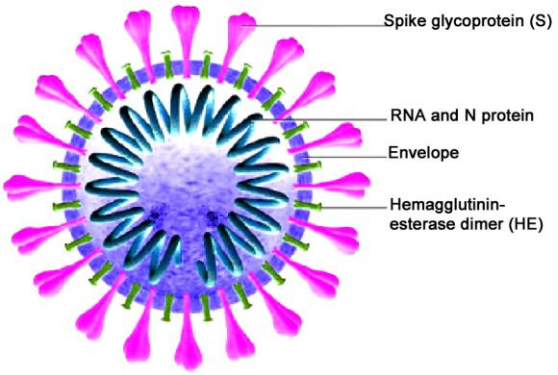

Figure 5. Structural depiction of coronavirus (Source: Drazen JM). Spike glycoprotein (surface glycoprotein) and the nucleocapsid protein sequences were analyzed in the current study. One 111 amino acid residues long fragment belonging to the surface glycoprotein is conserved in many lethal strains of human and bat coronaviruses as revealed by our analysis.

pairs. The program and its applications are more fully described in the methods section and by Figure 1. The validity of the program was established by usage that has reproduced results obtained by other programs with GeneBank data.

The nucleocapsid protein and surface glycoprotein of the SARS coronaviruses (Figure 5) are integral parts of the virus structure. They can be identical or can have varying degrees of similarity (homology) among viruses in this group (Table 1). As complex molecules on the virus surface, they are responsible for differences in host range, infectivity and pathogenicity.

The data presented strongly support a very close relationship among some bat and the novel human coronaviruses that are causing much morbidity and mortality across the world. The statistical probability of the occurrence of so many common permutations for two different proteins between any two strains of viruses purely by chance is infinitesimally small. The similarity between bat and human strains is, however, disputed by some scientists (8). With regard to this dispute and the natural selection hypothesis, we strongly disagree with the idea of using biological activities to determine the origins of viruses or any other species. We prefer, and have presented, physical evidence. Regardless of the dispute on the origins of the highly lethal virus strains, the high degree of homology (a physical characteristic) raises the theoretical possibility that vaccines against pre-2019 SARS strains will provide some cross-protection against the novel coronavirus strains. The comparative data between current and past strains of coronaviruses specifically establishes an approach for interim vaccine development. Avian, bovine, equine, canine, feline, calf-giraffe, rabbit, water deer, and some strains of human and bat coronaviruses have very low sequence homology with novel coronaviruses as analyzed by “Compare” (data not shown).

In conclusion, our data strongly support a close relationship among bat, the human SARS (2018 strain) and the novel coronavirus. The identified protein fragments are highly conserved in the lethal and in the highly-contagious SARS strains of the viruses including the older and the most recent ones. These proteins are essential to the virulence, lethality and infectivity of the viruses. They will be useful in designing vaccines and future improved diagnostic tests, and for understanding the nature of infection by these viruses and their potential future mutations.

6. REFERENCES

1. Deng SQ, Peng HJ. Characteristics of and public health responses to the coronavirus disease 2019 outbreak in China. J Clin Med. E575, 992(2020) DOI: 10.3390/jcm9020575 PMid:32093211 PMCid:PMC7074453

2. S. Vinga, J. Almeida: Alignment-free sequence comparison-a review.19, 513-23 92003) DOI: 10.1093/bioinformatics/btg005 PMid:12611807

3. L. Chen, W. Liu, Q. Zhang, K. Xu, G. Ye, W. Wu, Z. Sun, F. Liu, K. Wu, B. Zhong, Y. Mei, W. Zhang, Y. Chen, Y. Li, M. Shi, K. Lan, Y.Liu, RNA based mNGS approach identifies a novel human coronavirus from two pneumonia cases in

© 1996-2020

2019 Wuhan outbreak. Emerg Microbes Infect. 9, 313-319 (2020)
DOI: 10.1080/22221751.2020.1725399
PMid:32020836 PMCid:PMC7033720

4. F. Wu, S. Zhao, B. Yu, Y, M. Chen, W. Wang, Y.Hu, ZG. Song, ZW. Tao, JH. Tian, YY.Pei, ML.Yuan, YL. Zhang, FH. Dai, Y.Liu, QM. Wang, JJ. Zheng, L. Xu, EC. Holmes, YZ Zhang. A novel coronavirus associated a respiratory disease in Wuhan of Hubei province, China. National Center for Biotechnology Information. Submitted (17-JAN-2020)

5. EJ. Snijder, PJ. Bredenbeek, JC. Dobbe, V. Thiel, J. Ziebuhr, LL. Poon, Y. Guan, M. Rozanov, WJ. Spaan, AE. Gorbalenya. Biochem. Biophys. Res. Commun. 316, 476-483 (2004)

6. P. Zhou, XL. Yang, XG. Wang, B.Hu, L. Zhang, HR. Si, Y. Zhu, B. Li, CL. Huang, HD. Chen, J. Chen, Y. Luo, H. Guo, RD. Jiang, MQ. Liu, Y. Chen, XR. Shen, X. Wang, XS. Zhen, K. Zhao, QJ. Chen, F. Deng, LL. Liu, B. Yang, FX. Zhang, YY. Wang, GF. Xiao, ZL. Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 569, 270 - 273 (2020)
DOI: 10.1038/s41586-020-2012-7
PMid:32015507 PMCid:PMC7095418

7. Y. Tao, S. Tong. Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. Microbiol Resour Announc. 8, 00548 - 19 (2019)
DOI: 10.1128/MRA.00548-19
PMid:31296683 PMCid:PMC6624766

8. https://www.scienceboard.net/index.aspx?Sec= sup&Sub=can&Pag=dis&Itemid= 570

Abbreviations: NCP: Nucleocapsid protein, SGP: Surface glycoprotein, MERS: Middle East Respiratory Syndrome.

Key Words: Coronavirus, Nucleocapsid Protein, SARS0CoV-2, Sequence Homology, Surface Glycoprotein

Send correspondence to: Babu V. Bassa, Department of Environmental Toxicology 108 Fisher Hall, P.O. Box 9264 Southern University and A &M College Baton Rouge, LA 70813, Tel: 573-449-7444, Fax: 225/771-5350, E-mail: bbassa9824@gmail.com

© 1996-2020