Differences and Similarities between Coronavirus and other Viruses

Mohamed Abdul-Al[1], Raed Abd-Alhameed[1], Mansour Youseffi[1], Rami Qahwaji[1] and Simon J Shepherd[1]

{mabdulal@bradford.ac.uk, r.a.a.abd@bradford.ac.uk }

Faculty of Engineering and Informatics, University of Bradford, Bradford BD71DP, United Kingdom

Abstract. Coronavirus is the most dangerous virus in the world wide and it can easy spread between people, animals and plants because it is existing on one strand of RNA (Ribonucleic Acid) and it can duplicate faster than any virus. The source of coronavirus is still unknown, but some sources said that it came from seafood market and other sources said that it came from bat and snakes. It starts in Wuhan; China and every day the fatality increases. The symptoms are like a SARS-CoV (acute respiratory syndrome coronavirus)) and MERS-CoV (Middle East Respiratory Syndrome Coronavirus). By using nucleotide sequence of coronavirus from NCBI (National Center for Biotechnology Information) and some programs that ran on Matlab, the results show that there are some differences and similarities between coronavirus and other viruses such as Ebola, Flu-b, Hepatitis B, HIV and Zika especially for DEBs (distinct excluded blocks) program that shows at 5bp (base pair) there is a common with slightly difference between coronavirus "cgggg" and Ebola virus "cgtgg". The aim from this study is to find a way to help the doctors and scientists to stop spreading the coronavirus or to destroy it.

Keywords: Coronavirus (CoV), Ebola virus, HIV virus, Flu-b virus, Hepatitis B virus and Zika virus, DEBs

1 Introduction

A set of viruses that induce disorders in animals and humans are called coronaviruses. There are enveloped non-segmented positive sense of Ribonucleic acid (RNA) viruses associated to the family called Coronaviridae and the form Nidovirales and mostly appropriated in humans and other mammals [15]. Coronavirus has four subsets consists of:

i. Alpha-coronavirus
ii. Beta-coronavirus
iii. Gamma-coronavirus
iv. Delta-coronavirus

The genome size of these subsets is among 26 to 32 Kb (kilo-bases) [7, 10]. The 5' covered CoV genome contains of a 5' UTR (untranslated region), ORF (open reading frame), a 3' UTR and 3' poly(A) end. The 5' of 66.6% of the genome encode replicase related nonstructural

IMDC-SDSP 2020, June 28-30, Cyberspace

DOI 10.4108/eai.28-6-2020.2298247

proteins. While, the 33.3% of the genome encodes structural proteins. During replication in ill cells, a 3' coterminal nested a group of sub-genomic mRNAs is synthesized and these groups are also 5' coterminal with the director sequence of the genome [2, 11]. However, the first stage in CoV replication is synthesis of the negative strand RNA, and this mechanism is still hard to understand it [5].

In the past few years, there are 10,000 cases that had coronavirus (CoV) in humans and infections were mild the epidemics of the two beta-coronaviruses, serious SARS-CoV (acute respiratory syndrome coronavirus) [6, 8, 9] and MERS-CoV (Middle East Respiratory Syndrome Coronavirus) [1, 26], and it has death-rates 10% and 37% respectively [17, 23]. It was already recognized could only be the edge of the iceberg, with probably more unusual and serious zoonotic events to be report. Both SARS and MERS are identified by flu-like symptoms with fever, cough and anhelation and have the probability of transmission from animals to humans and vice versa or from one person to another [4, 12].

Recently, a coronavirus is a lethal disease that discovered in December 2019 in Wuhan, Hubei, China and spread too fast. There were a set of pneumonia cases with unknown cause (it likes a ghost), the diagnostic of this disease is the same as the diagnostic of respiratory infections [18]. 2019-nCoV (2019 novel coronavirus) was the name of the lethal disease after a deep sequencing analysis from the lower respiratory tract by World Health Organization (WHO). There were more than 800 cases that recognized especially in Wuhan in China and other cases in Japan, the USA, United Kingdom, South Korea and in Thailand [3, 19]. By 25^{th} of January 2020, there was 1975 cases has infected with coronavirus nationwide [14]. Also, in China mainland has seen hundreds of thousands of people left the city and carrying the virus with them. According to WHO, the source of coronavirus is still unknown. However, some studies stated that the source of coronavirus could be from bats or marmots or snakes that illegally sold wildlife in the market of seafood in Wuhan, China [24]. In this study, we are comparing CoV with other viruses to see the differences and similarities. Also, to find the reason why CoV is more dangerous than cancer. Moreover, to show for scientists and doctors from where they can start to stop or destroy CoV.

2 Methods

2.1 Data sources and searches

After a complete searching from China and worldwide official websites, prediction, advertisement and news [13, 14, 25]. The significant data that was collected from December 2019 till January 2020. Comparing coronavirus with other viruses.

2.2 Statistical Analysis

The data of genome codes of coronavirus and other viruses were collected from NCBI library, by using Matlab with different codes and analyzed them. In this study, five viruses such as Ebola, HIV, Flu-b, Hepatitis-B and Zika are compared with coronavirus in different codes such as gene sequence, probability and entropy, z-scores and distinct excluded blocks (DEBs) in genome codes.

3 Results

.

Table 1: Gene Sequence number for six viruses and word length.

Virus Name	Gene Sequence Number	Word Length(L)	Maximum	Minimum
Ebola	18960	L5	73 which is aaaaa	0 which is cgtgg
Flu-b	1841	L4	34 which is aaaa	0 which is cgta
Hepatitis B	3182	L4	32 which is cttt	0 which is cggt
HIV	9181	L5	52 which is aagaa	0 which is gtctt
Zika	10794	L5	51 which is tggag	0 which is taacg
Coronavirus	29903	L5	85 which is aaaaa	56 which is gaaaa
Coronavirus	29903	L8	26 which is aaaaaaaa	0 which is gaaaaaaa and taaaaaaa

Table 2: Entropy and Probability for six viruses.

Virus Name	Entropy	Probability of a	Probability of c	Probability of g	Probability of t	Total
Ebola	1.9736	0.3196	0.213	0.1981	0.2693	1
Flu-b	**1.9574**	0.3558	0.1917	0.2287	0.2238	1
Hepatitis B	1.9913	0.2297	0.2674	0.2175	0.2854	1
HIV	**1.9525**	0.3564	0.1788	0.2423	0.2224	0.9999
Zika	1.9864	0.2771	0.2185	0.2908	0.2135	0.9999
Coronavirus	**1.9570**	0.2994	0.1837	0.1961	0.3208	1

Table 3: Z-score for all viruses with string length (Two, Three and Four letters).

Virus Name	String Length (two, three and four letters)	Exact Number	Probability	Ratio	Z-Score
Flu-b	**aa**	195	**232.913**	0.8372	-0.3855
Hepatitis B	**tc**	281	**242.761**	1.1575	3.3916
Coronavirus	**aa**	2169	**2681.05**	0.809	-14.116

Ebola	**aaa**	415	**388.3737**	1.0686	1.8091
HIV	**aaa**	254	**201.015**	1.2636	4.9687
	gaa	283	**240.6727**	1.1759	3.7515
Zika	**aca**	229	**230.0653**	0.9954	-0.1028
	aga	280	**298.1201**	0.9392	-1.5176
Coronavirus	**aaa**	654	**525.4145**	1.2447	7.403
	caa	615	**490.0477**	1.255	7.37
Ebola	**taac**	157	**200.5874**	0.7827	-4.8231
Coronavirus	**taaa**	245	**229.4578**	1.0677	1.5237
	taac	176	**246.6496**	0.7136	-6.7914

Table 4: The difference between exact and total number for all viruses.

Viruses Name	Exact Number	Total Number
Ebola	17874	18960
Flu-b	1718	1841
Hepatitis B	2984	3182
HIV	8583	9181
Zika	10242	10794
Coronavirus	28206	29903

Table 5: Words that are not existing in all viruses.

Words that are not existing (dictionary)	Ebola	Flu-b	Hepatitis B	HIV	Zika	Coronavirus
4bp	0	10	2	0	0	0
5bp	1	307	133	69	12	1
6bp	286	2811	2129	1219	805	340

4 Discussion

In **gene sequence**, it can be observed how much each string length exists in the sequence of each virus. Also, it can be observed which string length has the maximum and minimum occurrence as shown in Table 1. For example, in all viruses except coronavirus, at L4 and L5 have maximum and minimum occurrence and the minimum is zero. While, in coronavirus at L5

still has 56 minimum occurrences such as ‘gaaaa’. In addition, by jumping from L5 to L8 it has two minimum occurrences are zero particularly ‘gaaaaaaa’ and ‘taaaaaaa’.

In Table 2 shows the **probability** of each letter in each virus and the total probability of each virus must be equal 1. However, for HIV and Zika viruses are approximately 1. In Ebola virus is a base pair (bp) because sometime if it is not bp it means random sequence. So, it is bp of 2 and the **entropy** is 1.97. The meaning of 2 is usually binary coded, identification to other viruses so sometimes can find the entropy is 1.95 such as Flu-b, HIV and CoV viruses that means coded but there are less than the binary code and there will never be greater than 2 such as Hepatitis B virus (approximately 2). If the entropy is 1.5 this means that something is missing in the coded. Entropy give us how far there is coded. Also, as it is seen at Table 2 the entropy of CoV (1.9570) is close to Flu-b virus (1.9574) and for HIV virus has (1.9524) which is smaller than CoV and Flu-b viruses. This means that CoV has some similar in nucleotide sequence to Flu-b virus. However, the length of nucleotide sequence for CoV is larger than Flu-b virus by 14 times.

The use of comparing two different scores that are from normal distribution and calculating the probability of a score taking place inside a normal distribution [16].
The biological significant of z-scores are words in genomes that are under and over represented. In Tables 3 shows that the **z-score** for all viruses with string length (two, three and four letters), exact number, probability (Noting that the probability should not be greater than 2), and ratio for each base pair in the sequence (Knowing that there are other than these string length but it is not possible to put it all). Furthermore, it is seen that the probability of Flu-b virus with string length especially **‘aa’** is **232.913**. In addition, in CoV with string length particularly **‘aa’** the probability of these string length is **2681.05**. Moreover, it is also like Hepatitis B such as **‘tc’**. In addition, for string length (four letters) the probability of Ebola virus ‘**taac’** has **200.5874**. While in CoV ‘**taac’** has **246.6496.** We can conclude that there is a common thing between these viruses in some code at nucleotide sequence and maybe it is difficult to destroy them even with a drug. If all the exact numbers are added, it will show that the exact number for each virus is less than the total number for each virus that is because there is some missing number which means that 2 letters are not exist in the sequence as shown in Table 4.

DEBs known as letter words that are not exist or not there in the sequence.
The significant of finding these DEBs in genome codes are to know how many base pairs (bps) in each virus are not exist in the sequence that means eliminate significantly of more words in the viruses than the number of words in the sequence. For example, in the Ebola, HIV and Zika viruses, the first DEB confronted is 5bp long in **Ebola** virus like **‘cgtgg’** and same as 6bp in **Ebola** (there are 286 words are not existing like **‘tttcgt’** and **‘tttgct’**) and above are not exist in the sequence. In addition, in **HIV** virus at 5bp long there are 69 words are not existing like **‘tgcgt’** and **‘ttcgt’**) and same as 6bp there are 1219 are not existing in the sequence. Moreover, in **Zika** virus at 5bp there are 12 words are not existing like **‘tatat’** and **‘tcggt’** and at 6bp there are 805 words are not existing in the sequence. Furthermore, in **CoV** at 5bp long like **‘cgggg’** is not existing as shown in Figure 1. While, at 6bp there are 340 are nor existing like **'aaagcg' and 'ttcgag'**. However, Flu-b and Hepatitis B viruses, the first DEB confronted is 4bp long in **Flu-b** (there are 10 words are not existing in the sequence like **‘ggcg’** and **‘tcgc’**) same as 5bp in **Hepatitis B** (307 there are 10 words are not existing in the sequence like **‘tttga’** and **‘tttgt’**) and above are not exist in the sequence as shown in Table 5, zero means that the words are existing in the sequence.

There is a slightly difference between Ebola virus **'cgtgg'** and CoV **'cgggg'** at 5bp that does not exists as shown in Table 10. This means that there is a change in coding process at the 5bp for CoV maybe it is decompressed. For example, if we image that c = 1, g = 0 and t = empty (-). In Ebola virus will be 10-00 and in CoV will be 10000. This implies the substrings from the compressed string and tries to replace the indexes with the corresponding entry in the anti-dictionary, which is empty at first and built up regularly.

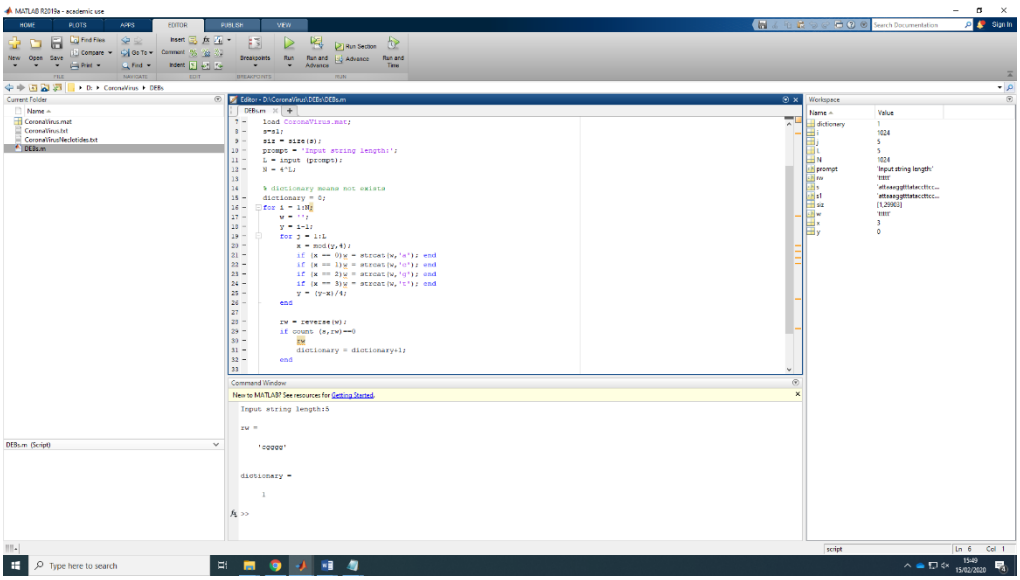

Fig.1. Shows the first not existing 5 letters in the gene sequence of 2019-nCoV with the program coding.

5 Conclusion

In this a paper, with these results, it is possible to control the 2019-nCoV after we found that in L5 'cgggg' is the place that 2019-nCoV can start to duplicate in the DNA and that is why it can spread faster than cancer. Therefore, from this point scientists and doctors can start working on this virus to stop it and find a treatment for all infected patients. However, they should represent that CoV is slightly same as Ebola virus with a different happened with coding process.

References

[1] Baker SC, de Groot RJ, Baric RS, et al. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol 2013; 87: 7790–92.

[2] Brian DA & Baric RS (2005) Coronavirus genome structure and replication. Curr Top Microbiol Immunol 287, 1–30.

[3] CDC. First travel-related case of 2019 novel coronavirus detected in United States. Jan 21, 2020. https://www.cdc.gov/media/ releases/2020/p0121-novel-coronavirus-travel-case.html (accessed Jan 23, 2020). [Online] [Accessed on February 08, 2020].

[4] Chen Y, Liu Q, Guo D. Coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020. [Epub ahead of print, 22 Jan 2020]. DOI:10.1002/jmv.25681.

[5] Chen-Yu Lo, Tsung-Lin Tsai, Chao-Nan Lin, Ching-Hung Lin and Hung-Yi Wu. (2019) 'Interaction of coronavirus nucleocapsid protein with the 50- and 30-ends of the coronavirus genome is involved in genome circularization and negative-strand RNA synthesis' Journal of the FEBS, pp.3222-3239.

[6] Drosten C, Günther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 2003; 348: 1967–76.

[7] ICTVK-AM (International Committee on Taxonomy of Viruses & King AM) (2012) Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses. Academic Press, London Waltham, MA.

[8] Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 2003; 348: 1953–66.

[9] Kuiken T, Fouchier RAM, Schutten M, et al. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. Lancet 2003; 362: 263–70.

[10] Lee S & Lee C (2014) Complete genome characterization of Korean Porcine deltacoronavirus strain KOR/KNU14-04/2014. Genome Announc 2, e01191.

[11] Masters PS (2006) The molecular biology of coronaviruses. Adv Virus Res 66, 193–292.

[12] Nature. Stop the Wuhan virus. Nature. 2020;577(7791):450.

[13] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information (NCBI). Nucleic Acids Res. 2016;44: D7–D19. https://www.ncbi.nlm.nih.gov/nuccore/NC_045512. [Online] [Accessed on February 06, 2020].

[14] NHC (National Health Commission's) briefing on the pneumonia epidemic situation. Released on 26 Jan 2020 (in Chinese). Available at http://www.nhc.gov.cn/yjb/s7860/202001/9614b05a8cac4ffabac10c4502fe517c.shtml [Online] [Accessed on February 08, 2020].

[15] Richman DD, Whitley RJ, Hayden FG, eds. Clinical virology, 4th edn. Washington: ASM Press, 2016.

[16] Virginia A., Afifi, Abdelmonem; May, Susanne K.; Clark, (2012), Practical Multivariate Analysis (Fifth ed.), Chapman & Hall/CRC, ISBN 978-1439816806.

[17] WHO (World Health Organization). Middle East respiratory syndrome coronavirus (MERS-CoV). November, 2019. http://www.who.int/emergencies/mers-cov/en/ (accessed Jan 19, 2020). [Online] [Accessed on February 07, 2020].

[18] WHO (World Health Organization). Novel coronavirus – China. Jan 12, 2020. http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (accessed Jan 19, 2020). [Online] [Accessed on February 07, 2020].

[19] WHO (World Health Organization). Novel coronavirus – Japan (ex-China). Jan 17, 2020. http://www.who.int/csr/don/17-january-2020-novel-coronavirus- japan-ex-china/en/ (accessed Jan 19, 2020). [Online] [Accessed on February 08, 2020].

[20] WHO (World Health Organization). Novel coronavirus – Republic of Korea (ex-China). Jan 21, 2020. http://www.who.int/csr/don/21-january-2020-novel- coronavirus-republic-of-korea-ex-china/en/ (accessed Jan 23, 2020). [Online] [Accessed on February 08, 2020].

[21] WHO (World Health Organization). Novel coronavirus – Thailand (ex-China). Jan 14, 2020. http://www.who.int/csr/don/14-january-2020-novel-coronavirus- thailand/en/ (accessed Jan 19, 2020). [Online] [Accessed on February 08, 2020].

[22] WHO (World Health Organization). Novel coronavirus – United Kingdom (ex-China). January, 2020 https://www.who.int/search?query=cases%20of%20coronavirus%20found%20in%20UK%20in%20january%202020&page=1&pagesize=10&sort=relevance&sortdir=desc&default=AND&f.Countries.size=100&f.Lang.filter=en&f.RegionalSites.size=100&f.Topics.filter=Coronavirus&f.Topics.size=100&f.contenttype.size=100&f.doctype.size=101&facet.field=RegionalSites&facet.field=Topics&facet.field=doctype&facet.field=Countries&facet.field=contenttype&facet.field=Lang&tune=true&tune.0=3&tune.1=2&tune.2=2&tune.3=3&tune.4=180&tune.5=75. [Online] [Accessed on February 08, 2020].

[23] WHO (World Health Organization). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Dec 31, 2003. https://www.who.int/csr/sars/country/table2004_04_21/en/ (accessed Jan 19, 2020). [Online] [Accessed on February 07, 2020].

[24] WHO. Statement on the meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). Released on 23 Jan 2020. Available at https://www.who.int/news-room/detail/23-01-2020-statement-on-the-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov). [Online] [Accessed on February 08, 2020].

[25] WMHC (Wuhan Municipal Health Commission's) briefing on the pneumonia epidemic situation. Released on 31 Dec 2019 (in Chinese). Available at http://wjw.wuhan.gov.cn/front/web/showDetail/2019123108989.

[26] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012; 367: 1814–20.