Characterization of Genome Sequence 2019 Novel Coronavirus (2019-nCoV) by using Bioinformatic Tool

N. A. Parmin¹, Uda Hashim¹,², N. Hamidah A Halim¹, M.N.A Uda³ and M.N.Afnan Uda¹,²

¹Institute of Nano Electronic Engineering, Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia.
²School of Microelectronic Engineering, Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia.
³School of Bioprocess Engineering, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia.

E-mail: azizahparmin@unimap.edu.my

Abstract. A novel coronavirus (2019-nCoV) became the seventh member in the family of Coronavirus that infect human. 2019-nCoV became the most severe virus compare to another family of coronavirus. Human airway epithelial cells have been used to identify and isolate the virus before proceed to reverse transcriptase Polymerase Chain Reaction (RT-PCR). Detailed biological knowledge is crucial for the development of effective countermeasures, diagnostic tests, vaccines and antiviral drugs against the 2019-nCoV. Conserved coding sequences within the spike glycoprotein region of open reading frame in the coronavirus genome was used as the basis to design oligonucleotide probe to detect the virus. Analyses on different strain of coronavirus sequences were done to check the percentage of similarity and consensus region that cause different strain of viruses. The biomarker needed an acceptable length between 22 and 31 mers. The choice of S gene region was identified and can be used as a biomarker probe for biosensor development. It has implications for coronavirus detection techniques in clinical and biosensor diagnostic system.

1. Introduction
A novel coronavirus (2019-nCoV) became the third pathogenic human coronavirus that emerged in the last two decades besides Severe Acute Respiratory Syndrome (SARS) coronavirus and Middle East Respiratory Syndrome (MERS) coronavirus [¹]. 2019-nCoV is a new member of coronavirus family that can cause serious respiratory diseases [²]. Coronaviruses are enveloped RNA viruses that can be transmitted among humans that cause severe acute respiratory diseases [³]. A novel coronavirus was detected in patients with pneumonia and the specimens were tested by the Chinese Center for Disease Control and Prevention (China CDC) at an early stage of the outbreak.

The results of independent sequencing of the coronavirus genome all indicated that it was a polyadenylated genomic RNA of 29.7 Kb in length [⁴,⁵]. Comparative analysis of the genome with other coronaviruses suggested that the virus genome was very similar to previously characterized coronaviruses, with the order (starting from the N-terminal): replicase (R), spike (S), envelope (E), membrane (M) and nucleocapsid (N) gene, where there are few accessory genes or motifs spanning...
between the structural genes and at the 3’UTR (untranslated region), which may not be necessary for viral replication.

To analyse the region of interest, BLAST programs are widely used as tools for looking protein and DNA databases for grouping similarities. BLAST have been figured to differentiate protein or DNA queries with databases, with DNA sequences frequently experiencing reasonable interpretation before any experimental procedures is performed. The aim of this research was to analyze the importance and usage of probe resides in coronaviruses region for sensing purposes through bioinformatics analysis such as BLAST and CLUSTALW.

2. Experimental

2.1 Analysis of 2019-nCoV Genome Sequences
The new genome sequence was obtained from GenBank by using FASTA sequences. Nucleotides sequences of different strains of coronavirus were identified; severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2 partial genome (MT198653.1), severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2 complete genome (MT226610.1), severe acute respiratory syndrome coronavirus 2 isolate nCoV-FIN, partial genome (MT020781.2), severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/NPL/61-TW/2020, complete genome (MT072688.1), and severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome (NC_045512.2). The sequence alignment produced by by MEGAHIT 1.2.9 (de novo assembly), using default parameters, was used to cross-validate with the reference-based method as an internal control. A reference genome in GenBank (accession number NC_045512) was used as a reference to compare with another strain of coronavirus. For phylogenetic analyses, coronavirus full-genome sequences were aligned with CLUSTAL W (6) using MEGA 10.0.5.[6,7].

2.2 Molecular characterization of the 2019-nCoV
2019-nCoV sequences can be utilized to investigate a sequence database by using a homology search of BLAST programming.

2.3 Phylogenetic relationship
Phylogenetic relationship was developed between 2019-nCoV and the other members of the same family through extensive comparison of their genome sequences.

3. Results and Discussions

3.1 Comparative Analysis of 2019-nCoV
Comparative analysis of the genome with other coronaviruses showed that 2019-nCoV virus genome was quite similar with previously characterized coronavirus; with the orientation order from the starting material of N-terminal, replicase (R), spike (S), envelope (E), membrane (M) and nucleocapsid (N) gene, where there are few accessory genes or motifs spanning between the structural genes and at the 3’ untranslated region (UTR), which may not be necessary for viral replication. Figure 1 showed the schematic diagram of 2019-nCoV including spike glycoprotein (S), membrane protein (M), Envelope, E protein, RNA and N protein.
The replicase gene encodes for two proteins (ORF1ab and ORF1a polyprotein) as a consequence of the proteolytic processing of the large protein as in table 1 [8]. The translation of two polyproteins (ORF1ab and ORF1a polyprotein) started the genome expression of 2019-nCoV. ATPase and DNA duplex-unwinding activity were demonstrated by purified helicase showed that the protein has RNA polymerase activity based on RNA genome sequences. The spike protein was identified to be incorporated with membrane protein into the viral envelope before the mature virion was released. The sequence of the gene product from ORF 3a shows no homology to any known proteins. The envelope protein of the 2019-nCoV was identified as the component of virus envelope [9]. The membrane protein (M gene) contained hydrophilic domain that believed to interact with the nucleocapsid protein. It was ligated inside the virus particle. M protein was identified can induce the immune response of human system. The ORF7a and ORF8 was showed for the existence of a cleavage site, same as ORF3a function. The N gene showed ORF9 structural protein for nucleosid phosphoprotein. The N gene sequence showed high similarity with the nucleocapsid protein of another coronaviruses.

The result showed that 2019-nCoV and another strain of coronaviruses had a similar genome organization but they acquired their accessory genes from RNA recombination with the infection host such as human and mammal. Their sequence variation implied their common ancestors, which might acquire from the RNA ligation into the host genome to disrupt their replication, transcription and translation mechanism.

Figure 2 showed multiple sequence alignment analysis by using ebi.ac.uk tools. Multiple sequence alignments analysis of the nucleotides residues on spike glycoprotein gene that produce surface glycoprotein of 2019-nCoV had been constructed. Columns indicate identical amino acids in the same position in all types analysed are exclusively indicated by an asterisk (*) which showed fully conserved residues, whereas alignments showing the presence of similar amino acids are represented by other symbols (:) for conservation of strong groups or (.) for conservation of weak groups. Symbol (-) showed no consensus of the groups.
The phylogenetic relationship between 2019-nCoV and another 3 types of coronaviruses has been constructed (figure 3). The evidence result by using phylogenetic tree studies was possible to identify the origin host of the coronavirus. The biological significance from this studies still remains to be elucidated.

Table 1. Features of 2019-nCoV Genome Sequence (NC_045512.2)

| Start-End | Gene   | Product                  |
|-----------|--------|--------------------------|
| 266-21555 | ORF1ab | ORF1ab polyprotein       |
| 266-13483 | ORF1ab | ORF1a polyprotein        |
| 21563-25384 | S  | Spike glycoprotein       |
| 25393-26220 | ORF3a | ORF3a protein            |
| 26245-26472 | E    | Envelope protein         |
| 26523-27191 | M    | Membrane protein         |
| 27202-27387 | ORF6  | ORF6 protein             |
| 27394-27759 | ORF7a | ORF7a protein            |
| 27894-28259 | ORF8  | ORF8 protein             |
| 28274-29533 | N, ORF9 | Structural protein       |
| 29558-29674 | ORF10a | Nucleosid phosphoprotein |

Figure 2. Multiple sequence alignment analysis by using
Figure 3. Phylogenetic tree schematic diagram of the relationship of all known coronaviruses based on the putative RNA recombination events occurred at the accessory ORFs.

The presence of the virus can be detected by using molecular approach such as Reverse Transcriptase Polymerase Chain Reaction (PCR). Measurable immune response rely on the analysis by using serological method such as enzyme-linked immunosorbent assay (ELISA).

Specificity of the PCR was enhanced by using amplification target of N gene region from genomic sequence of virus. N gene became the most abundant of subgenomic mRNA that produce during transcription process. Although the PCR was powerful device and most reliable method for detection of virus, the labor intensive and technically demanding became the disadvantages of the system. The development of biosensor technology for 2019-nCoV has capability of the rapid high screening and great future potential to be used as diagnostic tool.

4. Conclusions
The S gene had been used as a target gene for genotyping the coronaviruses strain such as human coronavirus compares to N gene and can be used as a designation of biomarker probe for the detection of 2019-nCoV virus. The S region was considered as the possible specific probe by comparison analysis using bioinformatic tools.

Acknowledgements
This research was supported by the Ministry of Higher Education (MOHE) under Grant Nos. FRGS/2/2014/TK03/UNIMAP/01/1 and the author would like to thank all staff members of the Institute of Nanoelectronic Engineering in Universiti Malaysia Perlis for their technical advice.

References
[1] Kampf G, Todt D, Pfaender S and Steinmann E 2020 J. Hosp. Infect. 104 246–251
[2] Wang C, Cheng Z and Yue X 2020 J. Risk Financ. Manag. 13 1–6
[3] N. Zhu 2020 *N. Engl. J. Med.* **382** 727–733
[4] Chow K Y C 2003 *Beijing Gen. Inst.* **1** 247–262
[5] Isolated S S 2020 March **2** 19–21
[6] Wisecaver J H and Hackett J D 2014 *Mol. Phylogenet. Evol.* **71** 184-192
[7] Parmin N A, Hashim U and Gopinath S C B 2018 *Int. J. Biol. Macromol.* **107** 1738-1746
[8] Harcourt B H 2004 *J. Virol.* **78** 13600-13612
[9] Ramaiah A and Arumugaswami V 2020 *BioRxiv.* **29** 925867