Microbes and Infectious Diseases 2022; 3(4): 780-783

Microbes and Infectious Diseases

Journal homepage: https://mid.journals.ekb.eg/

Short report

Large-scale analysis of SARS-CoV-2 envelope protein sequences reveals universally conserved residues

Vivek Darapaneni

Department of virology and computational biochemistry, Anvek Institute of Biomolecular Research, Visakhapatnam, India.

ARTICLE INFO

Article history:
Received 3 July 2022
Received in revised form 23 July 2022
Accepted 25 July 2022

Keywords:
SARS-CoV-2
Envelope protein
Viroporin
Conserved
COVID-19

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the ongoing COVID-19 pandemic that has devastated mankind with an unprecedented impact on both health and economic condition globally. The envelope protein of SARS-CoV-2 is a multifunctional viroporin across endoplasmic reticulum-Golgi intermediate compartment. SARS-CoV-2 envelope (E) protein plays a crucial role in the virus life cycle. The objective of the present study was to identify the residue conservation in the SARS-CoV-2 E protein. The study was based on 2,654,250 amino acid sequences for the E protein. On the whole, this study exposed residues that are universally conserved among different strains of SARS-CoV-2. These universally conserved residues might be involved in either structure stabilizing or protein-protein interactions. The conserved residues identified in the present study in conjunction with structural analysis of the E protein could form the basis for designing universal anti-SARS-CoV-2 drugs which are resistant to mutations arising in the future.

Introduction

Coronaviruses are enveloped viruses with a large positive strand RNA genome. Coronaviruses have been known to infect humans, bats, civets, cattle, horses, swine, dogs, cats, turkeys, rabbits, chickens, rats and mice. Coronaviruses cause gastroenteritis and respiratory tract diseases in their hosts. Previously HKU1, NL63, 229E and OC43 coronaviruses have been known to cause mild respiratory disease in humans. In the year 2002, a novel coronavirus has crossed over from bats to humans through palm civet cats [intermediary host]. Again a decade later in 2012, another novel coronavirus crossed over from bats to humans through dromedary camels [intermediary host]. The former virus was named severe acute respiratory syndrome coronaviruses (SARS-CoV), while the later virus was named middle east respiratory syndrome coronaviruses (MERS-CoV). In the year 2019, yet again a novel coronavirus had crossed over from bat to humans via malaya pangolin [intermediary host]. This virus was designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 causes severe respiratory tract infection and causes a global pandemic resulting in deaths worldwide. As of 23 June, 2022, approximately 539 million confirmed cases of SARS-CoV-2 infection were reported globally. Of which approximately 6.3 million deaths were reported globally [1].

SARS-CoV-2 genome encodes four main structural proteins and sixteen non-structural proteins [2]. Coronavirus protein expression in the infected cell occurs in two different and distinguishable phases. In the first phase, the 5'-

DOI: 10.21608/MID.2022.148868.1340
* *Corresponding author:* Vivek Darapaneni
E-mail address: vivek110385@gmail.com

terminal region (≈ two thirds) of the viral genome is translated into sixteen non-structural proteins, NSP1–NSP16. In the second phase, the 3'-terminal region (≈ one third) of the viral genome is translated into four structural proteins namely, spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins [2].

The SARS-CoV-2 E protein is a single pass transmembrane protein with three distinct domains namely, the N-terminal domain (residues 1-7), transmembrane domain (residues 8-38) and C-terminal domain (residues 39-75) [3]. The E protein of SARS-CoV-2 is a multifunctional viroporin across endoplasmic reticulum- Golgi intermediate compartment and plays an important role in virus assembly, virion release and viral pathogenesis [4]. M and E protein interaction are required for viral particle formation [5]. The E protein residues Tyr2, Phe26-Thr30, Arg38, Leu39, Tyr42, Lys63 and Leu74 are involved in M and E protein interaction [6]. The E protein of SARS-CoV-2 contains a PDZ binding motif between the residues 72 to 75 [7].

In this study, we determined the evolutionary conserved and variable regions in SARS-CoV-2 E protein across all strains of SARS-CoV-2. This study will provide a basis for designing universal mutation resistant drugs for future evolving strains of SARS-CoV-2.

Method

SARS-CoV-2 E protein sequences were retrieved from National Centre for Biotechnology Information (NCBI) SARS-CoV-2 resource (https://www.ncbi. nlm.nih.gov/sars-cov-2/), belonging to all pangolin strains of SARS-CoV-2. Protein sequences with ambiguous characters were discarded. Duplicate sequences were removed using CD-HIT web server (http:// -lab. ucsd.edu/cdhit_suite/ cgi-bin/ index.cgi) [8]. Multiple sequence alignment of the obtained protein sequences was performed using MAFFT version 7 (https://mafft. cbrc.jp/alignment/ software/) [9]. Full length structure of E protein was predicted by the Robetta server using the RoseTTAFold method (https://robetta. bakerlab.org/submit.php) [10].

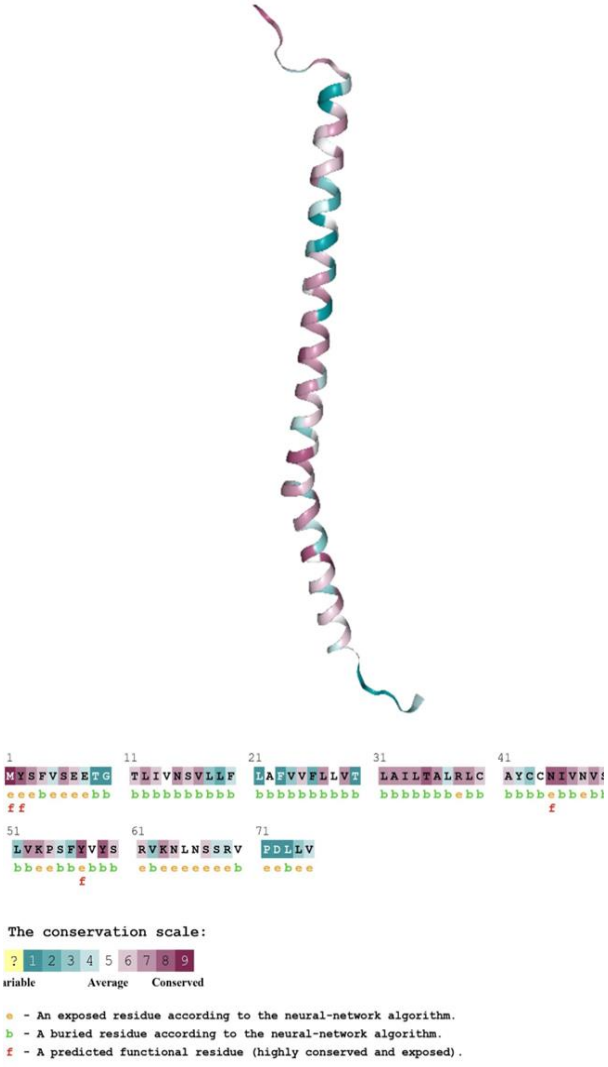
RoseTTAFold uses a three-track network, that is successful transformation and integration of information at the sequence level (one dimensional), distance map level (two dimensional) and coordinate level (three dimensional) [10]. The confidence score given by the server ranges from 0 to 1. "0" means a bad model whereas "1" means a good model [10]. By giving E protein structure and multiple sequence alignment files as an input, conserved and variable regions in E protein were identified using the ConSurf server. (https://consurf.tau.ac.il/) [11]. ConSurf algorithm produces meaningful conservation scores by taking evolutionary relationships among protein sequences into account. The conservation score given by the ConSurf server is divided into nine grades. The variable positions in the protein are placed in grade 1 to grade 4, intermediately conserved positions are placed in grade 5, and the conserved positions are placed in grade 6 to grade 9 [11].

Results and Discussion

A total of 2,654,250 SARS-CoV-2 E protein sequences were retrieved. After the removal of identical sequences, a total of 667 unique sequences was obtained. Multiple sequence alignment of the obtained sequences was carried out using MAFFT version 7. Full length structure of E protein was obtained from the Robetta server with a confidence score of 0.87 which suggests that the obtained model is highly reliable. Conservation analysis of E protein was carried out using the consurf server by giving E protein structure as well as multiple sequence alignment file as input and is illustrated in **figure (1)**.

The conserved residues detected in the SARS-CoV-2 E protein may have either functional importance or structural importance. On the contrary, variable sites arise as a result of either adaptation or evolutionary pressure to evade the host immune system. The N-terminal domain of E protein showed high sequence conservation. The transmembrane domain and C-terminal domain of E protein showed intermediary conservation. The conserved residue positions Thr11 to Val17 and Leu31 to Ala36 might play a pivotal role in the ion channel activity of E protein. The conserved residues Tyr2, Leu27, Leu28, Val29, Arg38, Leu39 and Lys63 play a critical role in the interaction between E and M proteins. The PDZ binding motif of E protein showed high degree of variation in the region suggesting that PDZ binding is not universally conserved among different strains of SARS-CoV-2.

Figure 1. The residue conservation view of the SARS-CoV-2 E protein obtained by projecting conservation scores onto the protein structure and sequence.



Conclusion

In culmination, this study has identified that SARS-CoV-2 E protein showed a pattern of conserved and variable residues among all strains of SARS-CoV-2. Identifying drug binding sites near the conserved residues in the E protein, it will help in developing anti-SARS-CoV-2 drugs which are unlikely to get ineffective in case of mutation of SARS-CoV-2 into drug resistant form. Moreover the anti-SARS-CoV-2 drugs targeting these binding sites are universally efficient against SARS-CoV-2. Further work arising from this study should characterize the function of the previously unknown highly conserved residues.

Conflict of interest

The author report no conflict of interest.

Financial disclousure: None.

References

1-**WHO.** WHO coronavirus disease (COVID-19). dashboard. Available at: https://covid19.who.int/. Accessed June 23 2022.

2-**V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V.** Coronavirus biology and replication: implications for SARS-CoV-2. Nature Reviews Microbiology 2021;19(3):155-70.

3-**Mandala VS, McKay MJ, Shcherbakov AA, Dregni AJ, Kolocouris A, Hong M.** Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. Nature structural & molecular biology 2020;27(12):1202-8.

4-**Schoeman D, Fielding BC.** Coronavirus envelope protein: current knowledge. Virology journal 2019; 16(1):1-22.

5-**Huang Y, Yang ZY, Kong WP, Nabel GJ.** Generation of synthetic severe acute respiratory syndrome coronavirus pseudoparticles: implications for assembly and vaccine production. Journal of virology 2004;78(22):12557-65.

6-**Mahtarin R, Islam S, Islam MJ, Ullah MO, Ali MA, Halim MA.** Structure and dynamics of membrane protein in SARS-CoV-2. Journal of Biomolecular Structure and Dynamics 2022; 40(10):4725-38.

7-**Teoh KT, Siu YL, Chan WL, Schlüter MA, Liu CJ, Peiris JM et al.** The SARS coronavirus E protein interacts with PALS1 and alters tight junction formation and epithelial morphogenesis. Molecular biology of the cell 2010;21(22):3838-52.

8-**Huang Y, Niu B, Gao Y, Fu L, Li W.** CD-HIT Suite: a web server for clustering and

comparing biological sequences. Bioinformatics 2010;26(5):680-2.

9- **Katoh K, Rozewicki J, Yamada KD.** MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in bioinformatics. 2019;20(4):1160-6.

10- **Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al.** Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021; 373(6557):871-6.

11- **Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al.** ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic acids research 2016;44(W1):W344-50.

Darapaneni V. Large-scale analysis of SARS-CoV-2 envelope protein sequences reveals universally conserved residues. Microbes Infect Dis 2022; 3(4): 780-783.