In Silico Domain Structural Model Analysis of Coronavirus ORF1ab Polyprotein

Mohammed I. Jameel, Rabar J. Noori and Soma F. Rasul

Department of Medical Microbiology, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region – F.R. Iraq

***Abstract*—The world today is battling with a coronavirus infection that is considered a global pandemic. Coronavirus infection is mainly attribute to the varying technique of the replication and release of different genomic components of the virus. The present study aims to establish the physical and chemical features, as well as the basic structural and functional properties of *Coronavirus ORF1ab* domain. A molecular approach was adopt in this study using the Swiss Model and Phyre2 server whereas the prediction of the active ligand binding sites was done using Phyre2. The analysis of the structure of the protein showed that it has good structural and heat stability, as well as better hydrophilic features and acidic in nature. Based on the Homology modeling, only two binding active sites were noted with catalytic function being mediated by Zn^{2+} as the metallic heterogeneous ligand for binding sites prediction. The proteins mostly exhibited helical secondary configurations. This study can help in predicting and understanding the role of this domain protein in active coronavirus infection.**

***Index Terms*—Coronavirus, Ligand, Modeling, ORF1ab.**

I. Introduction

Coronavirus disease 2019 (COVID-19) is a public health problem facing the world today; it is an infectious disease caused by the coronavirus family which has wreaked havoc to both human lives and the global economy since its emergence in December 2019. It is a global pandemic that has infected over 50 million and killed nearly 1,300,000 persons (World Health Organization, 2020). The disease is more severe among elderly people who have accounted for the largest population of the affected cases (Dowd, et al., 2020; Wu, 2020; Onder, 2020). Coronavirus 2019 has been confirmed in more than 200 countries globally and new cases are being reported in other regions. With the pattern of its transmission, it is believed that a vaccine is urgently needed to develop strong immunity that is needed to prevent a relapse of the disease in the future (Kisslear, et al., 2020).

ARO-The Scientific Journal of Koya University
Vol. X, No. 2 (2022), Article ID: ARO.10829. 4 pages
DOI: 10.14500/aro.10829
Received: 10 June 2021; Accepted: 28 July 2022
Regular research paper: Published: 25 August 2022
Corresponding author's e-mail: mohammed.isam@koyauniversity.org

Statement deleted. Coronavirus (due to COVID-19 occur in 2019 [Schaecher and Pekosz, 2010; McBride and Fielding, 2012]) is an RNA virus (positive stranded) with structural proteins that include envelope protein (E), spike protein (S), membrane protein (M), and nucleocapsid phosphoprotein (N). In addition, the structural genes encode nine accessory proteins, encoded by ORF3a, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF14, and ORF10 genes (Nelson, et al.). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic (Schaecher and Pekosz, 2010; McBride and Fielding, 2012). The genomic attributes of the novel coronavirus are believed to contribute to the ongoing pandemic (Wu, et al*.,* 2020; Paraskevis, et al., 2020); the envelope protein ion channel (CoV EIC) is responsible for tweaking discharge of virion (Li, et al*.,* 2020). Other severe acute respiratory syndrome SARS 1 and 2 like coronaviruses also have spike proteins and ORF8 and ORF3a proteins but their pathogenicity and pattern of transmission differ from those of SARS-CoV (To, et al*.*, 2020).

Studies have found that the spike proteins of the novel coronavirus facilitate its penetration of epithelial cells through surface interaction with angiotensin-converting enzyme 2 (ACE2), thereby causing human infection. The role of ORF8 and ORF3 in viral infection remains unknown due to the limitations of the existing experimental techniques. However, the mechanism of infection of the severe acute respiratory syndrome-2 is still enigmatic (Rothe, et al*.,* 2020). Following the identification of the structures of the viral proteins of SARS-CoV-2 in early February 2020, many studies have been dedicated to the discovery of novel medicine that may be effective in treating SARS-CoV-2 infection (Muhammed, et al*.,* 2020). The aim of this study to establish the physical and chemical features of coronavirus using different tools.

II. Methods

A. Recovery of the Amino Acid Sequence of the Polyprotein

The polyprotein A.A. sequence was use in the present study was retrieve from the official NCBI sequence databank accessible at https://www.ncbi.nlm.nih.gov. The sequence was sourced with the Accession No: ORF-1; Sequence number: MN908947.3, Gene – ORF1; Polyprotein QHD43415.1; Polyprotein - ORF1; Organism – Pneumonia virus; identifier – Wuhan-Hu-1.

Physiochemical Characterization

Different web servers were used to determine the physiochemical characteristics of the protein sequence. For instance, the structure, aliphatic indices, elimination (coefficient), and instability of the amino acids were predicted using the ProtParam tool (Expasy), coupled with the prediction of the isoelectrical point (pI) and grand average of hydropathicity (GRAVY) (Gasteiger, et al., 2005).

B. The Prediction of Secondary Structure of Protein

The secondary structure of the polyproteins was predicted using the PROFsec, PSIPRED, and SOPMA programming tools (Geourjon, et al., 1995) whereas the DISOPRED tool was used for disorder prediction (Rost, et al., 2004).

C. Prediction of Polyprotein Binding Sites and Gene Ontology

The profISIS server was used or the prediction of the protein-protein binding sites (Ofran and Rost, 2007); this was done through identification of the interacting moieties from the sequence alone through the consolidation of the predicted structural features with the developmental information. The gene ontology (GO) prediction approach was also employed for the prediction of the molecular, cellular, and biological characteristics using the homology to known annotated proteins (Hamp, et al., 2013).

D. Homology Modeling and Polyprotein Validation

Being that the + (3-D) structure in the protein data bank cannot be accessed, this unexplored space domain was modeled using 700 a. a. long-protein sequence. Hence, two programs (Swiss-Model) and (Phyre-2) were used to perform the protein homology modeling (Waterhouse, et al., 2018). The prediction of the secondary protein structure was done using Phyre2 (Kelly, et al., 2015) whereas the Swiss-Model was used to produce the 3D protein model. After the analysis, only the most reasonable (3D) model was selected for the validation step. The Ramachandran plot analysis was used for the approval of the last modeled structure whereas PROCHECK was used for the analysis of the stereochemical property. Finally, the modeled structure was uploaded to the 3D Ligand Site web server for potential binding site prediction (Wass, et al., 2010).

III. Results and Discussion

A. Physiological and Chemical Characterization of the Polyprotein

Coronavirus amino acid sequence was obtained as FASTA file with an instability index (ii) of 27.52 and used as a query sequence for the physiochemical characterization. The protein was acidic in nature (+ve = 76, -ve = 77) whereas the Pi was 7; the MW of the protein was 79,822.08. The protein exhibited a high elimination coefficient value of 102,245 which suggests the existence of +ve Cys residues. However, the aliphatic index value of 77.690 of the search protein implies a good level of stability over a temperature range. As shown in Table I, the protein exhibited a low GRAVY indices value of –0.204 which suggests a hydrophilic nature that creates a better room for reaction with water (Kyte and Doolottle, 1982).

B. Prediction of Secondary Structure of Polyprotein

The employed default boundaries (window width – 17; similarity threshold – 8; division factor – 3, and number of states – 4) were used for the prediction of the secondary protein structure with the SOPMA tool. With the aid of 511 proteins (used as sub-database) and seven aligned proteins, the following predictions were made using the SOPMA tool: Irregular coils = 35.29% of residues, alpha-helix =38.43% of residues, beta-turn = 9% of residue, and extended strand = 17.29% of residue (Table II). This suggests that the protein has a higher chance of having a helix, strand, and coil, as shown in Fig. 1. The prediction of the secondary protein structure by (PROFsec) (Predict-Protein) using neural organization tool showed the protein to have a prediction precision of >28.71% for helix configuration (α; π; 3_10- helix), 21.43% beta-strand+, and 49.86% loop (L). DISOPRED-based intrinsic disorder profile processing showed that most of the amino acid (>95%) did not meet the recommended +0.5 certainty confidence score for disordered condition which is the least chance for bending, thereby suggesting that the predicted protein is of high strength (Fig. 2).

C. Prediction of Gene Ontology and Binding Sites of Protein

The profISIS software was used for protein binding sites prediction; 14 different binding sites were identified at positions 2; 114–115; 160; 211–212; 249; 316; 417; 482–483;

TABLE I
Physiochemical Properties of the SARS-CoV-2 Polyprotein Using the ProtParam Tool

S/n.	Attribute	Value
1	Number of amino acids	700.0
2	Molecular weight	79,822.080
3	Isoelectric point	7.0
4	Instability index	27.520
5	Aliphatic index	77.690
6	Extinction coefficient	102,245.0
7	Total number of –vely charged amino acids	77.0
8	Total no. of +vely charged amino acids	76.0
9	GRAVY (Grand average of hydropathicity)	–0.2040
10	Chemical formula	$C_{3586}H_{5457}N_{947}O_{1038}S_{43}$

TABLE II
Secondary Protein Structure Prediction Using SOPMA Tool

Secondary protein structure element	Value (%)
Alpha helix (Hh)	38.430
310 helix (Gg)	0.000
Pi helix (Ii)	0.000
Beta bridge (Bb)	0.000
Extended strand (Ee)	17.290
β-turn (Tt)	9
Different coil (Cc)	35.290

512; 560; 584; 647; 696; and 700 whereas gene ontology (GO) predicted and classified the utilitarian viewpoints as cellular, molecular, and biological. Catalytic activity includes nucleotide transferase and transferase activity whereas molecular function (MF) includes restriction which involves heterocyclic, organic, and cyclic compound binding, as well as small molecule and nucleic acid restricting activity. From GO analysis, we found that the high outcome in catalytic activity and RNA binding with more than 50. On the other hand, we found the last outcomes in RNA polymerase and transferase activity equal to 41. Whereas the same functions were also observed in the case of biological process (BP) but all functions have sane score equal to (61) and in case of cellular component (CC), all functions have score equal to 63 and in host cell membrane, the score was 54.

D. Modeling of Protein Homology and Validation of X-domain Structure

The sequence of the target protein was keyed in as the information file in the workspace of the employed Swiss-Model. A search in the Swiss-MODEL layout library (SMTL) using HHBlits yielded around complete 50 formats. Out of the 50 templates, only 6 nur.A1 was the suitable template. Then, the objective sequence was selected based on the outcome of the Qualitative Model Energy Analysis (QMEAN) (-0.77), identity sequence proportion (95.83), global model quality estimate (GMQE) 0.74, and coverage (93%). The model relied on track format alignment using ProMod3 with remodeling of the Indel and revamping of the side chains. The achieved model in this study was similar to 6nur.1. The generated model was saved in the PDB format, as shown in Fig. 3. The assessment of the stereochemical nature of the predicted protein structure was done using the Ramachandran plotting map, as shown in Fig. 4. The assessment showed that 534 of the all-out residues (91.1%) domiciled in favored area (A; B; and L) whereas 9% domiciled in the permitted area (a; b; and l). No residue was found in the prohibited region. The quality of the X domain model was predicted through analysis of 118 good resolution structures (2.0 A°) with R-factor of <20%. Deviation of 5.8 in maximum residue properties, as well as bond angle of 5.3 and 93.1% planar, was observed within the cutoff point through PROCHECK analysis. PHYRE2-based modeling of protein homology showed a protein model that modeled 77% of the protein residues (538) at a confidence level of >90% (Fig. 5) which

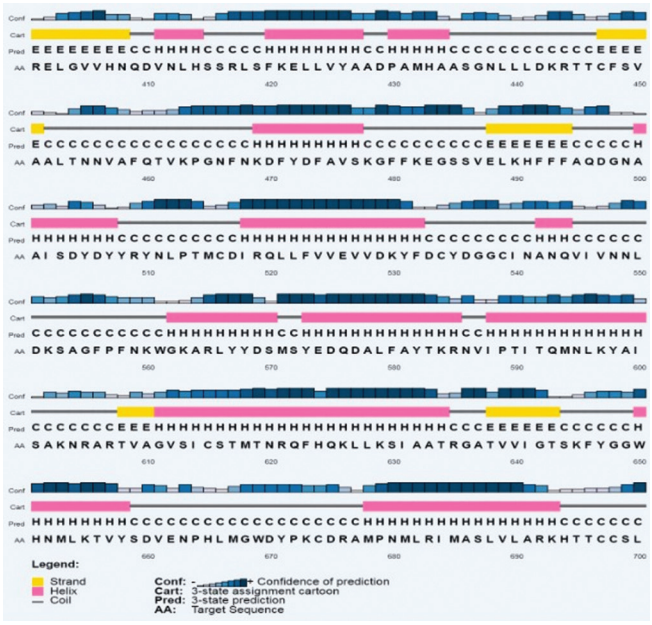

Fig. 1. PSIPRED-based 2° protein model prediction.

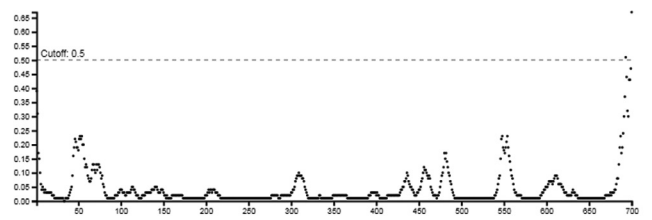

Fig. 2. Observed randomness condition from the DISOPRED Web.

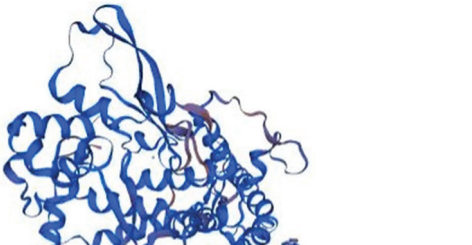

Fig. 3. SWISS-Model-predicted protein structure with helix strands and coil.

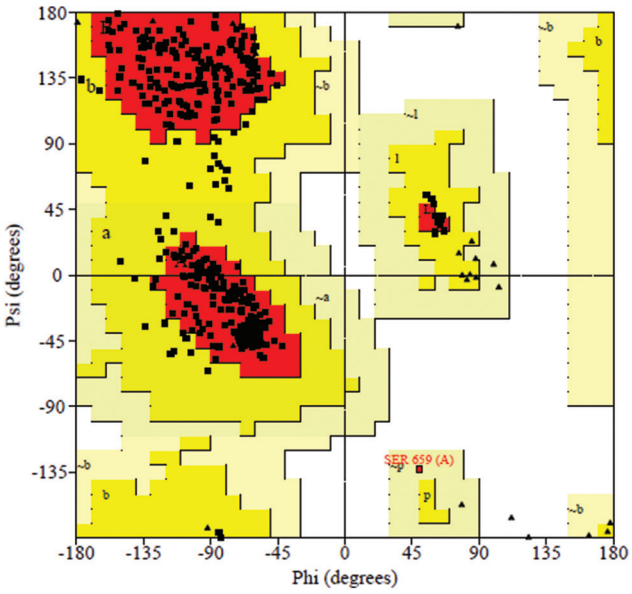

Fig. 4. Ramachandran plot of the structure of the predicted protein.

http://dx.doi.org/10.14500/aro.10829

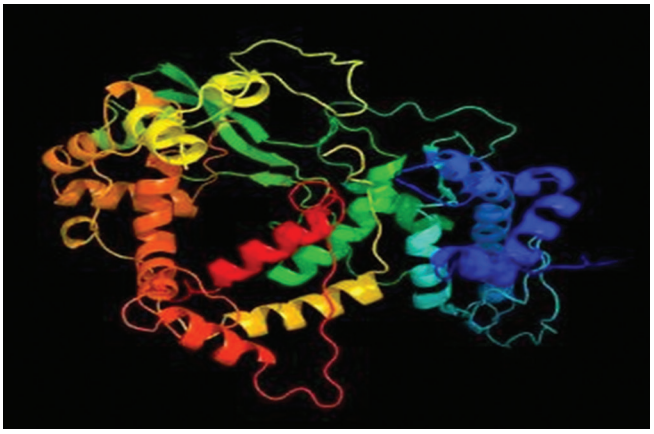

Fig. 5. PHYRE2-predicted protein structure.

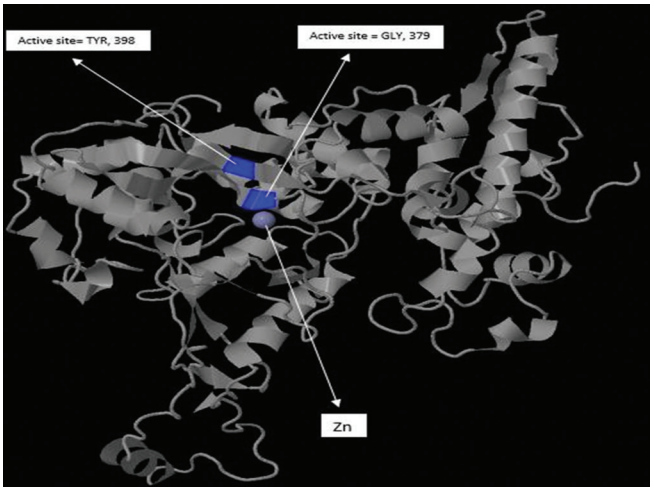

Fig. 6. 3D protein model plus Zn^{2+} ligand on the binding site.

implies the following coordinates X = 72.1690; Y = 75.6500; and Z = 86.0880. The expected secondary structure by Phyre2 was given as 53% alpha-helix, 13% disordered, and 11% beta strand. In the final selected model, the prediction of the 3D ligand site was reliant on Cluster 1 which showed 1 ligand and 1 structure. Only two binding sites were successfully predicted as follows Gly: 1 contact (at residue no. 379) and Tyr: 1 contact (at residue no. 398). The predicted metallic heterogeneous ligand was Zn (1) which binds to the binding site of the predicted ligand (Fig. 6).

Conclusion

We determine the domain structural model of coronavirus orf1ab polyprotein with predicted active site for ligand binding site. This supplies premeditations into the functional role of polyprotein domain in viral pathogenesis.

References

Gasteiger, J. and Walker, M., 2005. *The Proteomics Protocols Handbook*. Springer, Berlin, p.571.

Geourjon, C. and Deleage, G.C., 1995. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Computer Applied Bioscience*, 681(11), pp.681-684.

Gil, S.C. and Von Hippel, P.H., 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Analytical Biochemistry*, 182(2), pp.319-326.

Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hönigschmid, P., Hopf, T.A., Kaufmann, S.A., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M. and Rost, B., 2013. Homology-based inference sets the bar high for protein function prediction. *BMC Journal Bioinformatics*, 14(7), pp.502-511.

Kelly, L.A, Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6), p.845.

Kyte, J. and Doolottle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1), pp.105-132.

Li, S., Yuan, L., Dai, G., Chen, R.A., Liu, D.X. and Fung, T.S., 2020. Regulation of the ER stress response by the Ion channel activity of the infectious bronchitis coronavirus envelope protein modulates virion release, apoptosis, viral fitness, and pathogenesis. *Frontiers in Microbiology*, 10, p.3022.

McBride, R. and Fielding, B.C., 2012. The role of severe acute respiratory syndrome (SARS) coronavirus accessory proteins in virus pathogenesis. *Viruses*, 4(11), pp.2902-2923.

Ofran, Y. and Rost, B., 2007. ISIS: Interaction sites identified from sequence. *Bioinformatics Journal*, 23(2), pp.e13-e16.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G. and Tsiodras, S., 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection Genetics and Evolution*, 79, pp.104-212.

Raik, G. and Luis, S., 2010. Strategies for protein synthetic biology. *Nucleic Acids Research Journal*, 38(8), pp.2663-2675.

Remmert, M., Andreas, B., Andreas, H. and Johannes, S., 2012. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), pp.173-178.

Rost, B., Guy, Y. and Liu, J., 2004. The predicProtein server. *Nucleic Acid Research Journal*, 32(2), pp.321-326.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., Seilmaier, M., Drosten, C., Vollmar, P., Zwirglmaier, K., Zange, S., Wölfel, R. and Hoelscher, M., 2020. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10), pp.970-971.

Schaecher, S.R. and Pekosz, A., 2010. *Molecular Biology of the SARS-Coronavirus*. Springer, Berlin, pp.153-166.

To, K.K.W., Tsang, O.T., Yip, C.C., Chan, K.H., Wu, T.C., Chan, J.M., Leung, W.S., Chik, T.S., Choi, C.Y., Kandamby, D.H., Lung, D.C., Tam, A.R., Poon, R.W., Fung, A.Y., Hung, I.F., Cheng, V.C., Chan, J.F. and Yuen, K.Y., 2020. Consistent detection of 2019 novel coronavirus in saliva. *Clinical Infectious Diseases*, 71(15), pp.841-843.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G. and Jiang, T., 2020. Genome composition and divergence of the novel coronavirus (2019- nCoV) originating in China. *Cell Host and Microbe*, 27(3), pp.325-328.

Wu, F., Zhao, S., Yu, B., Chen, Y., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), pp.265-269.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F. and Tan, W., 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8), pp.727-733.