Comparative Domain-fold analysis of the SARS-CoV-2 ORF1ab polyprotein: Insight into co-evolution, conservation of folding patterns, potential therapeutic strategies, and the possibility of reemergence

Srijeeb Karmakar[1]; Sachin Kumar[1], Vimal Katiyar[2]

[1]Department of Bioscience and Bioengineering,

[2]Department of Chemical Engineering,

Indian Institute of Technology Guwahati,

Guwahati, Assam-781039

Email address**:** karma176106011@iitg.ac.in, sachinku@iitg.ac.in, vkatiyar@iitg.ac.in

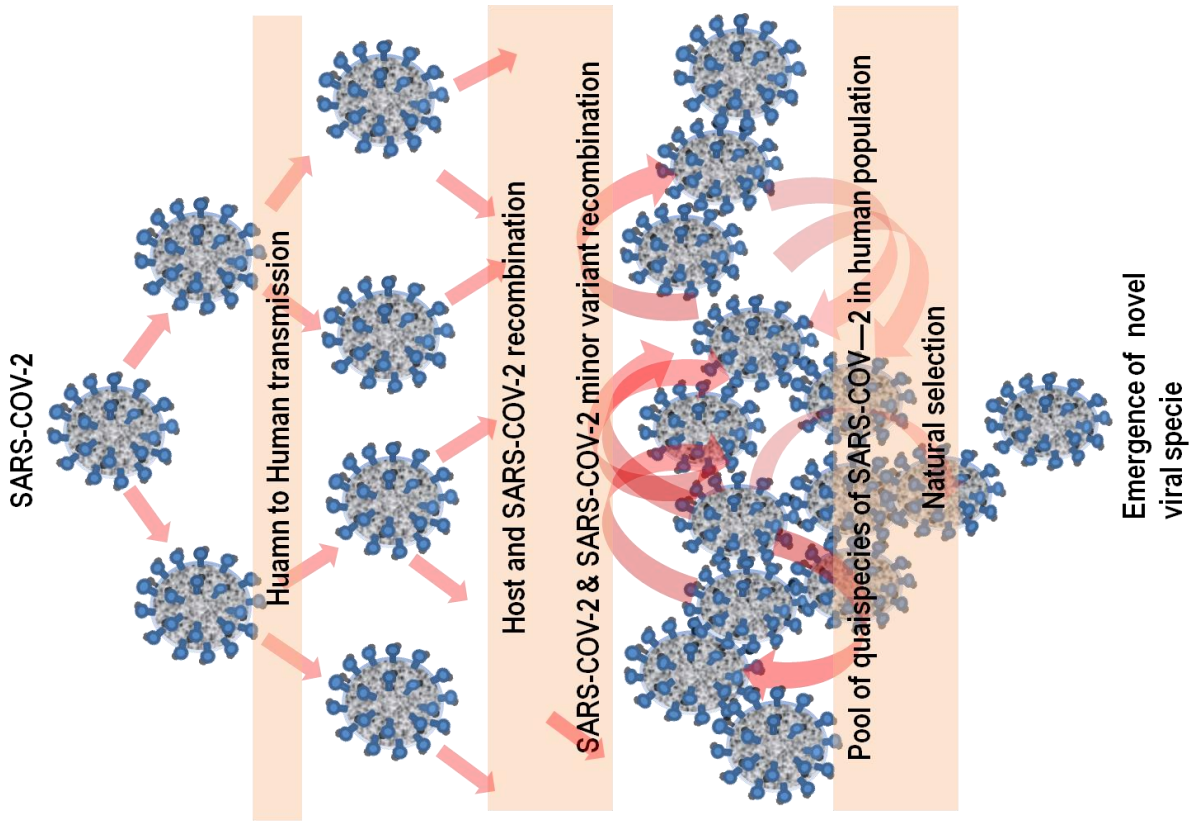
Abstract

The high transmissibility and replication of SARS-CoV-2 have been attributed to enhanced protein functions which are dependent on protein folding. Our *in silico* study endeavored to scrutinize SARS-CoV-2 ORF1ab by analyzing the conserved folding patterns of its transcribed proteins. Accordingly, the findings indicated that SARS-CoV-2 ORF1ab shares domain-specific fold-fingerprints with a spectrum of unrelated organisms. Closer observation revealed slight changes in folding patterns engendered with small variation in the intrinsic amino acid sequence. By correlating with the evolvability-potential of RNA-viruses and COVID-19 pandemic, we hypothesize that SARS-CoV-2 could undergo fast recombination with the host, SARS-CoV-2 minor variants and other viral species resulting in a reservoir of SARS-CoV-2 quasispecies. It is highly possible that natural selection will cause a future emergence of evolved SARS-CoV-2-descendants. Nonetheless, we hope that this insightful study will assist in elucidating SARS-CoV-2 protein functionalities, development of vaccines, and the possibility and nature of future emergence.

Keywords: SARS-CoV-2; domain-fold; evolutionary fingerprint; conserved domain; point mutation; recombination

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

Graphical Abstract



Highlights

- ORF1ab polypeptides contain conserved domain-specific folding patterns shared with unrelated organisms
- Minor variation in folding is fostered with minor change in sequence.
- COVID-19 pandemic might cause creation of large pool of SARS-CoV-2 minor variants
- SARS-CoV-2 high transmissibility and replication rate might accumulate variations at a fast rate.
- SARS-CoV-2 might evolve through recombination, quasispecies pool, and natural selection
- Possibility of future outbreak should be considered strongly.

1. Introduction

Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) has been reported to exhibit enhanced protein functionalities and transmissibility, leading to an unprecedented rise of a worldwide pandemic in the present times (Walls et al., 2020; Weber et al., 2020; Yang

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

and Chen, 2020). As of April, 2020, the pathogen has infected more than million individuals since December, 2019, killing thousands. The aggravation of the COVID-19 pandemic has been attributed to enhanced protein functionalities which are crucial for pathogenesis and high replication rate (Liu et al., 2020). For example, the affinity of the SARS-CoV-2-Spike protein with ACE2 receptors which is pivotal for host cell entry has been found to be enhanced in SARS-CoV-2 in comparison to SARS-CoV (Yan et al., 2020). In order to find therapeutic strategy, it is therefore the crucial and urgent need of the hour to elucidate SARS-CoV-2 evolutionary mechanisms (Tang et al., 2020) through which the pathogen integrates advantageous information in its genome and augments protein function.

In this context, it should be reflected that conformational domains of proteins are conserved more than the amino acid sequences to retain functionality (Illergård et al., 2009; Zhang et al., 2004), i.e, the folded topography of protein domains are of greater evolutionary significance. However, the folding mechanisms have also been found to be dictated by the intrinsic amino acid sequences (Dobson, 2003; Ma et al., 2007). In order to discern the evolutionary mechanisms of SARS-CoV-2, this study endeavours to draw the nexus between the aforementioned concepts, and attempts in interpreting the effect of small changes in the intrinsic amino acid sequence, causing slight changes in folding mechanisms, and thereby, engendering evolution of protein function.

The SARS-CoV-2 open reading frame, ORF1ab, which encodes for the Non-structural proteins (NSPs), was chosen for the study as it is vital in the replication of SARS-CoV-2 genome (Wu et al., 2020). The NSPs are quintessential contributors to the formation of SARS-CoV-2 replication machinery, bolstering RNA transcription, and thereby, the replication of the SARs-CoV-2 inside the host cell (Qiu and Xu, 2020). The NSPs are also involved in collapsing the innate immunity of the host (Jain et al., 2020), which is crucial in the success of SARS-CoV-2 in causing the exponential rise of COVID-19 fatalities. Therefore, it was been strongly suggested that the NSPs of ORF1ab are the potential targets for developing vaccines and novel drugs (Jain et al., 2020; Qiu and Xu, 2020), which requires the elucidation of their three dimensional folding characteristics.

Therefore, this study scrutinized evolutionary fingerprints of SARS-CoV-2 ORF1ab dictating the folding mechanisms of its transcribed proteins. The intrinsic amino acid sequences of the ORF1ab was thoroughly analysed employing simple computational tools of the Phyre2 fold recognition server, developed by Kelley LA *et al.* at the Imperial College, London (Kelley et

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

al., 2015). The aforementioned server was used because it links the amino acid sequence of a protein with a library of three dimensional domain-folds across all species. Correspondingly, the ORF1ab transcription products were compared with an elaborate fold library to understand conserved domains, sequence alignment, structural variations, and the possible evolutionary route. The study thereby attempts to frame a hypothesis based upon the SARS-CoV-2 conserved domains, and the possibility of co-evolution of SARS-CoV-2 ancestors infecting host cells of different species to utilize them as recombination platforms, and integrate evolutionary advantageous genetic information to enhance protein functions. As the viral lifecycle depends entirely on the host cell, achieving improved adaptation to the host cells must have been adopted as a strategy for a stronger survival. Moreover, experimental reports about the fast genomic recombination of SARS-CoV-2 is already emerging (Yi, 2020).

The study further takes an unconventional approach to form an assumption that the hypothesis is true, and explores the possibility of future re-emergence (Li et al., 2020), taking into account COVID-19 infections occurring across geographical, multiethnic, multi-cultural environments, with the possible creation of multiple recombination platforms, and a large pool of SARS-CoV-2 quasispecies. Shen et al., while studying the genomic diversity of SARS-CoV-2 in the COVID-19 patients, have also emphasized on similar lines (Neher et al., 2020; Shen et al., 2020).

2. Materials and methods

Simple computational tools were employed to analyze the multiple proteins of the SARS-CoV-2 ORF1ab. Amino acid sequences of the ORF1ab proteins were derived from Wuhan-hu-1 isolate (Accession id. NC_ 045512.2)(Zhu et al., 2020) . The FASTA sequences were submitted to the Phyre2 fold recognition server (http://www.sbg.bio.ic.ac.uk/phyre2) which utilizes a multi-component platform to search homologous proteins, predict three dimensional structure, align unrelated sequences to predict structural features, and also predict the fold-similarity between protein domains (Kelley et al., 2015). The server uses HHblits to determine evolutionary profile by capturing residue preferences at each position along the length of the query sequence which is scanned against a sequence-database. To obtain the secondary structure of the query sequence, the program PSIRED is used. The aforesaid program uses neural networks for the prediction of α-helices, β-strands and coils with an average three-state accuracy of 75–80% (Remmert et al., 2012). The sequence profile and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

the secondary structure predicted are converted into a Hidden Markov Model (HMM). The HMM obtained is then scanned across a fold library, a pre-compiled database of experimentally verified protein conformations, to search for the best HMM-HMM matches. The detection method uses an algorithm named as HHsearch (Söding, 2004), which generates a list of query-template alignments which are ranked according to posterior probabilities. Homology modelling is carried out by matching against a library of 2-15 amino acid-fragments of known protein structures, and subsequently, a sequence-profile search is conducted to find best matches for the missing sequences. Finally the fragments are fitted into a crude model using cyclic co-ordinate descent (Canutescu and Dunbrack, 2003). The complete modelling is carried out by multiple template modelling with Poing, a protein folding simulator. It creates a complete model even when different domains of the protein sequence have been modelled by different templates or without a template (Jefferys et al., 2010). The backbone of the protein model is reconstructed using Pulchra (Rotkiewicz and Skolnick, 2008).

The results generated with the Phyre2 fold recognition server was subsequently scrutinized manually, to check for simple changes in the intrinsic amino acid sequence and subsequent structural change between the predicted structure of the query sequence and the known structure of templates. The domain analysis was interpreted with the rationale that the proteins with maximum domain coverage were evolutionarily closely related, and vice versa.

3. Results and discussion

3.1 SARS-CoV-2-NSP1 (Leader protein)

The NSP1 leader protein of the SARS-CoV-2 is predicted to resemble the SARS-CoV NSP1-like fold very strongly with 88% domain-fold similarity from the residues 13-127. The likelihood of a common folding pathway appears to be high between these proteins, and the possibility of exhibiting similar structural domains is most probable. Accordingly the SARS-CoV-2 leader protein has been interpreted to attain a conformation with 42% alpha helix, 13% beta strand and 19% disordered region, with predominant features like hydrogen bonded turns, bends and 3_{10} helices. However, if the dissimilarity is scrutinized, it could be attributed to the change in the single amino acids in the sequence from SARS-CoV to SARS-CoV-2. For example, the amino acid residue at the positions (a) 84/85 and (b) 92/93 of the SARS-CoV-2- NSP1 was observed to be variant from the otherwise aligned sequence of the SARS-CoV- NSP1 [(a)VM to KV; (a) LE to MD)], shown in Figure 1A. It was found that the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

residue change increased the propensity of alpha helix-beta strand transition which is crucial in changing folding pattern of domains, thereby enhancing or modifying protein function (Ménade et al., 2018). The physiological impact of such transition of alpha helix/beta sheet transition had been suggested by Karmakar et al.(Karmakar et al., 2018). If such conformational change is engendered with a slight change of intrinsic amino acid sequence through recombination by integrating host cell sequence (de Haan et al., 2008), it would lead to one of the biggest challenges to deal with SARS-CoV-2 pandemic. The presumption of the study is that the pandemic outbreak of the virus infecting diverse human populations will increase the possibility of point mutations, giving rise to massive numbers of conformational variants of crucial proteins. Amongst such a large pool of conformational variants, it is alarmingly possible that proteins involved in the pathogenesis of SARS-CoV-2 will enhance or modify protein function in the natural select.

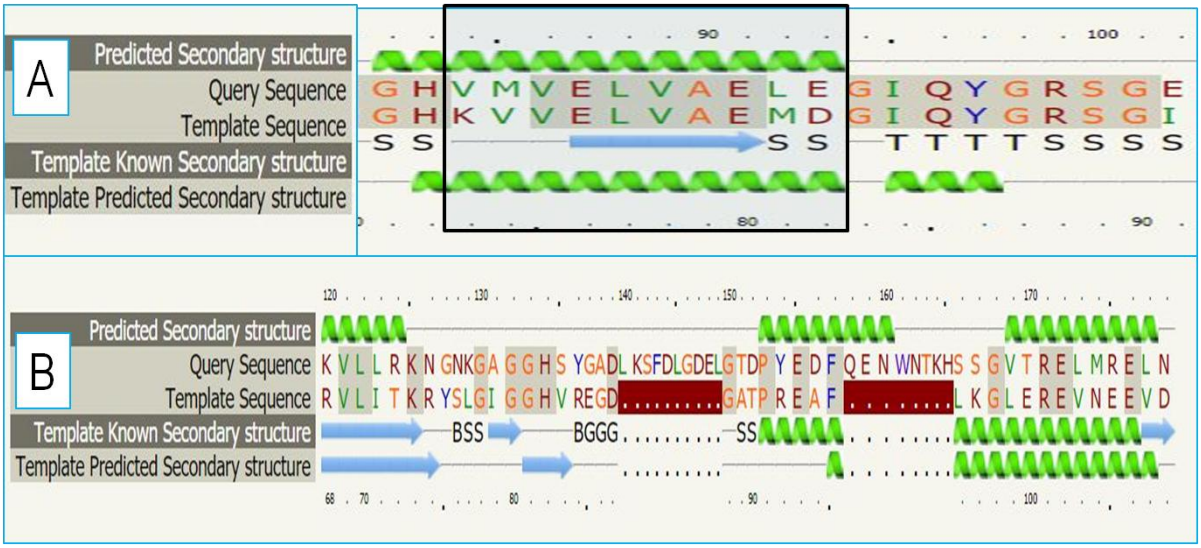


Figure 1. (A) Variation in sequence alignment at position 84/85 and 92/93 increases the propensity of secondary structure transition between SARS-CoV-2-NSP1 and SARS-CoV-NSP1; (B) doman-specific Sequence alignment between SARS-CoV-2-NSP1 and TM1382 (putative nudix hydrolase, PDB: 3E57) of *Thermotoga maritima* (ATCC:43589).

Interestingly, SARS-CoV-2-NSP1 reserves an intrinsic evolutionary fingerprint of domain folding similar to a protein,tm1382, of the hyperthermophillic organism, *Thermotoga maritima* (Huber et al., 1986) from the residues 120-178 (Shown in Figure 1B). This is speculative of the evolutionary pathway of SARS-CoV-2 integrating thermophillic characteristics and therefore, its ability to quickly adapt to extreme conditions is indeed a possibility worth considering. Here, we hold a strong view that extreme geographical and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

cultural jumps of SARS-CoV-2 might integrate drastic and novel information into the organism (Neher et al., 2020), and even trigger SARS-CoV-2 to re-express evolutionary signals discarded in the present form of the organism.

3.2 SARS-CoV-2-NSP2

The SARS-CoV-2-NSP2 appears to be a novel protein and exhibits extremely low similarity to any of the known conformations in the fold library. The domain analysis indicates large evolutionary divergence of the protein as significant similarity in sequence alignment and domain conformations were not found. This is suggestive of the distinct conformational profile of SARS-CoV-2-NSP2 generated by incorporation of folding patterns from a diverse pool of distant proteins. Therefore, it is suggested that the conformational nature of the protein is highly chimeric, with the possibility of distinct and enhanced function, which presumably contributes to the pathogenesis and replication of SARS-CoV-2. Figure 2A illustrates the chimeric nature of the SARS-CoV-2-NSP2 exhibiting conserved domains in scattered oligopeptide stretches from a large pool of species (Figure generated by Phyre2 fold recognition server).

Nevertheless, the protein appears to exhibit a domain conformation similar to the E3 Ubiquitin-protein ligase hectd1 from the residues 257-312, with 27% similarity in domain-conformation. The residues 143-165 appears to have an evolutionary fingerprint shared with the signaling peptide, RAS guanyl-releasing protein 1. Furthermore, the domain fold stretching from the residues 305-331 appears to have an imprint, although with significantly less shared-identity, of the viral polyproteins found in human and bovine enteroviruses (Shown in Figure 2B). Such diverse shared identities in domain-folds indicate the flow of genomic information to SARS-CoV-2 from humans and other species which might have been integrated through the process of recombination (Fang et al., 2005). It is also anticipated here that the entire proteome of SARS-CoV-2 has the possibility of exhibiting shared conformational identity with crucial proteins of humans and other infective viruses, the clarity of which might unfold a resolution to combat the spread of the COVID-19 pandemic. This is yet again emphasized that this must be investigated thoroughly whether the SARS-CoV-2 is evolving gradually through infection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

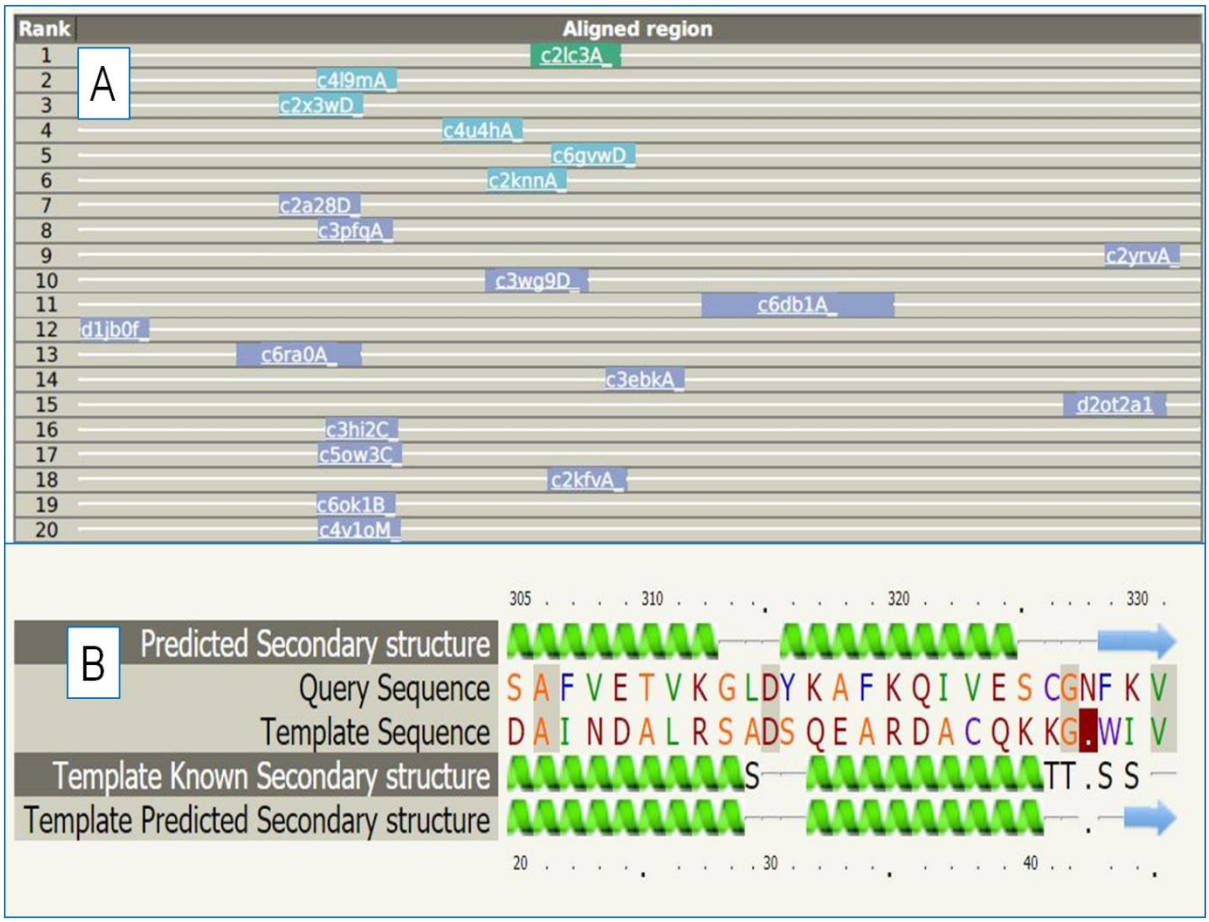


Figure 2. (A) Chimeric nature of the SARS-CoV-2-NSP2 exhibiting conserved domains in scattered oligopeptide stretches from a large pool of species; (B) Conserved domain stretching from the residues 305-331 shared with 'Crystal structure of human ACBD3 domain in complex with 3A protein of enterovirus-D68 (fusion protein, LVVY mutant)'[PDB: 6HMV]

3.3 SARS-CoV-2-NSP3

Similar to SARS-CoV-2-NSP2, NSP3 was also predicted to exhibit distinct structural novelty and evolutionarily appears to be largely divergent according to the domain-fold analysis report (Shown in Figure 3A). The domain folds, as predicted, does not display significant coverage to any existing protein conformations, however, shares identity with domains from a diverse spectrum of unrelated proteins. The native conformation of SARS-CoV-2-NSP3 is predicted to fold in a manner exhibiting 38% alpha helix, 25% beta strand, 21% disordered region and 8% TM helix. Residues 675-1059 appears to share folded conformation similar to diverse proteins: tandemly linked domains of nsp3 from murine hepatitis virus, papain-like

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

protease from MERS-CoV, avian infectious bronchitis virus, and SARS-CoV papain-like protease with maximum identity similarity of 82%.

Also, 2-111 residues of the SARS-CoV-2-NSP3 possess the potential of folding into a domain recognized in SARS-CoV with 77% shared identity. Furthermore, the domain-fold most likely attained by the residues 1089–1203 appears to have an evolutionary fingerprint with RNA-binding protein of the SARS-CoV. Residues 212- 374 appears to conserve a domain fold which contains imprints from evolutionary divergent proteins like human Poly[ADP-Ribose] polymerase 9, ADP-Ribose-binding protein from the *Oryza sativa*, *Streptomyces coelicolor's* macrodomain protein (chain-A). Most strikingly, the domain also appears to share structural identity, although low, to APPR-1-P processing domain of the thermophillic organism, *Thermus aquaticus*. It further provokes a speculation that SARS-CoV-2 might have thermophillic imprints across its proteome which has been acquired through overlapping evolutionary pathways. However miniscule, this aspect, as stated earlier, must be taken into consideration as fast evolution of the SARS-CoV-2 to re-express thermophillic attributes would be challenging to deal with. Yet another conformational prediction of SARS-CoV-2-NSP3 is its structural component suggesting the presence of a multi-domain transmembrane helix with both the N and C terminals embedded inside the cytoplasmic space. Therefore, it is verily emphasized that SARS-CoV-2-NSP3 exhibits a novel functionality which contributes strongly to the life cycle of SARS-CoV-2, making it recalcitrant to therapeutic action. Nevertheless, the protein also shares identity with domains present in other viral species associated with curable infectious diseases which opens the possibility that the library of compounds with the potential of blocking those predicted domains might also show action against SARS-CoV-2-NSP3.

The salient feature of SARS-CoV-2-NSP3 was predicted to be a transmembrane protein with six subunits (S1-S6) which is shown in Figure 3B. The N-terminal and the C-terminal both are embedded withdrawn inside the viral body. Therefore, it is presumable that the signaling cascade is conducted by the domains in the residues 1365-1379, 1440-1490 and 1551-1555. The insets of Figure 3B displays the structural features of these domains.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

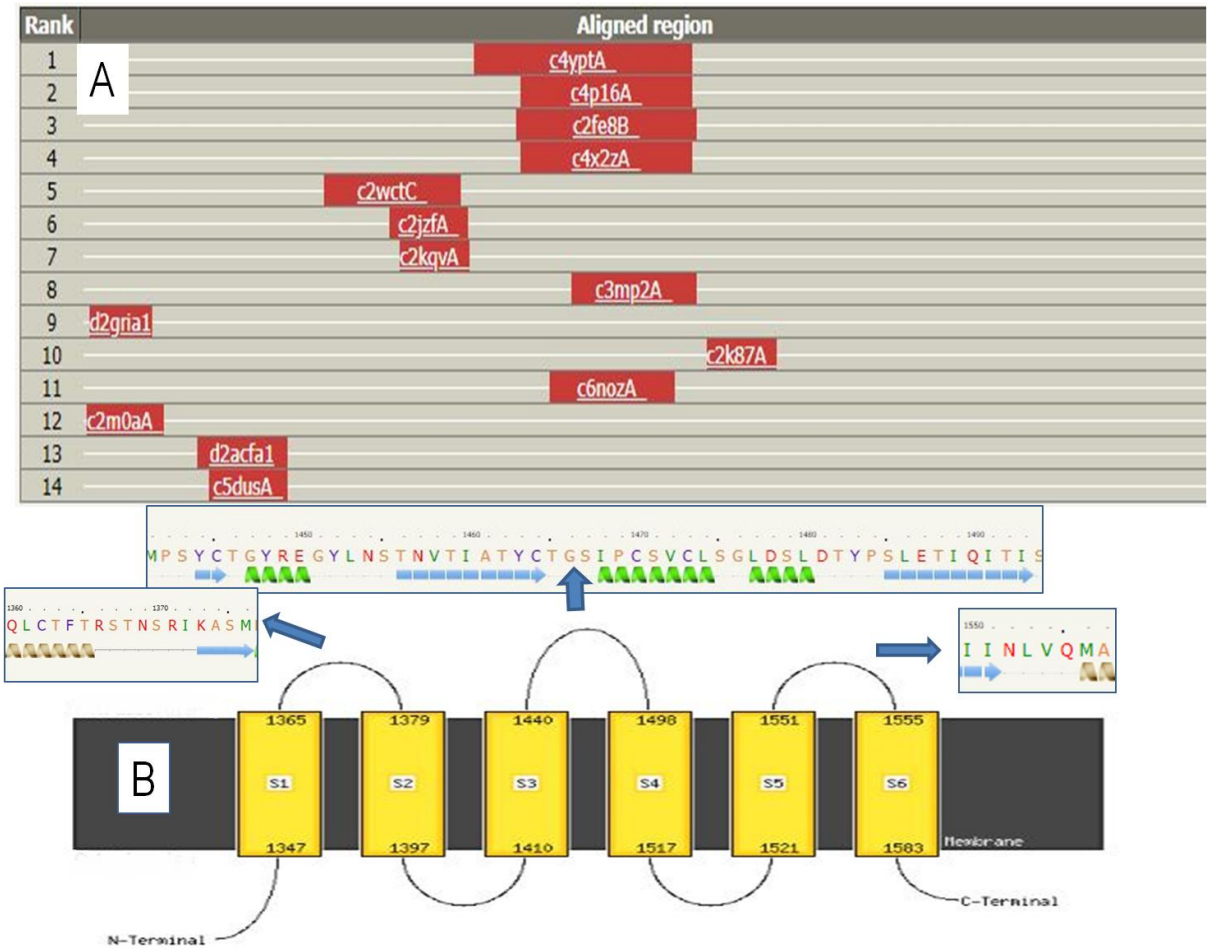


Figure 3. (A) Chimeric nature of the SARS-CoV-2-NSP3 exhibiting conserved domains in scattered oligopeptide stretches from a large pool of species; (B) The predicted topography of SARS-CoV-2-NSP3 transmembrane subunits. The inset shows the structural features of the residues forming the outer protrusions of SARS-CoV-2-NSP3.

3.4 SARS-CoV-2-NSP4

The SARS-CoV-2-NSP4 protein also appears to be a greatly distinct and evolutionarily divergent, with two distinct oligopeptide conserved domain (shown in Figure 4A). However, the most striking characteristic that was revealed in its evolutionary profile was a shared domain identity with several polymorphs of human alpha defensins which are associated with innate immunity. The residues that constitute the defensin-like fold extend from the positions 217 to 237, and exist as the most conserved domain of the protein. The figure shows the domain analysis report of SARS-CoV-2-NSP4 which shows that a minimum of 18 proteins, mostly belonging to the class of alpha defensin, share domain identity with the aforementioned NSP4 stretch. Indeed it is intriguing that a protein of SARS-CoV-2 possesses

a conserved domain like the human defensins involved in innate immunity. Moreover, it is well observed that the COVID-19 fatality is increased with decreased personal and herd immunity (Porcheddu et al., 2020). Therefore, it is strongly emphasized here that the aforementioned observation must be scrutinized in precise details to elucidate the nexus in between.

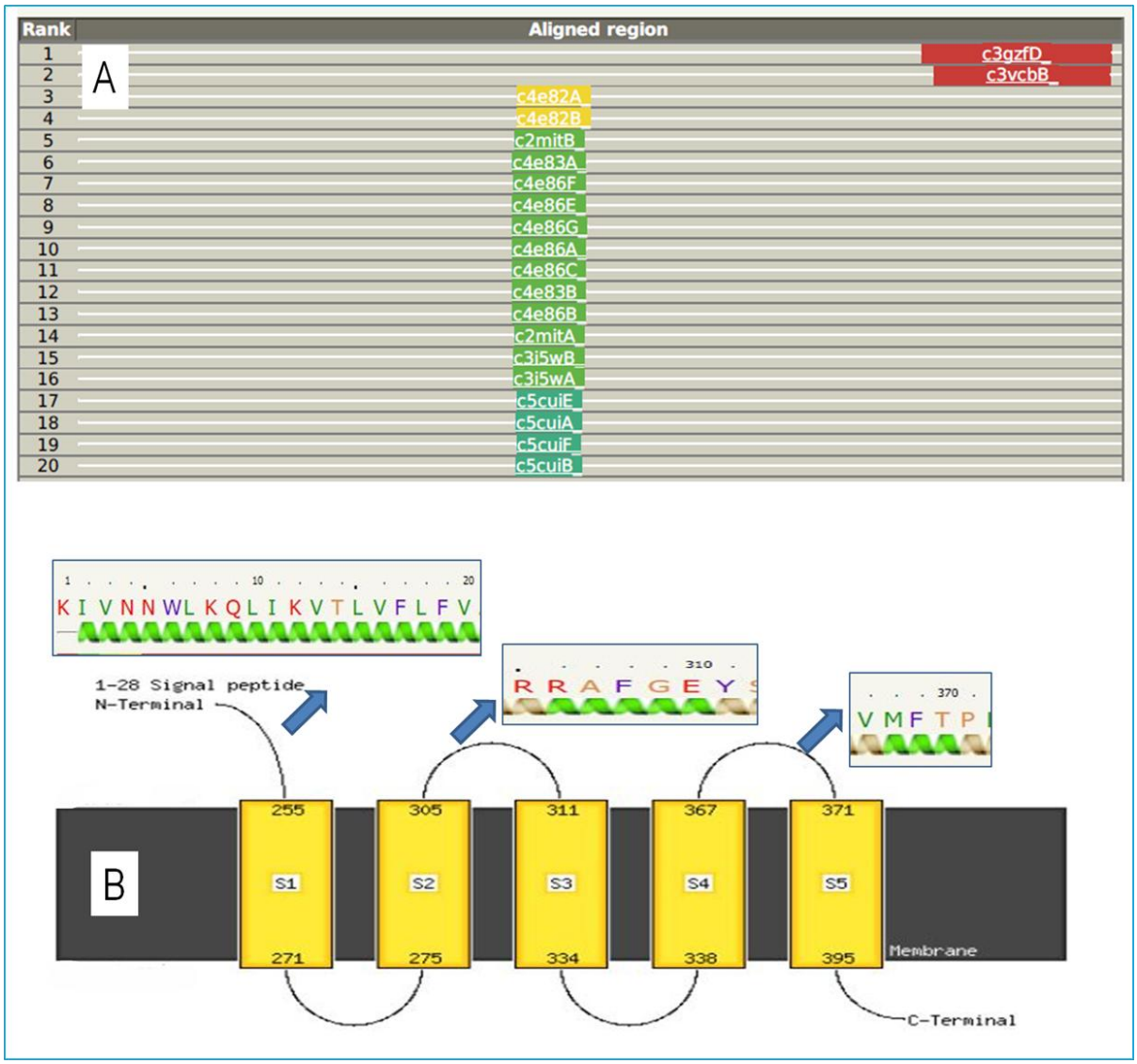


Figure 4. (A) Domain analysis report of SARS-CoV-2-NSP4 protein showing two distinct oligopeptide conserved domain; (B) Transmembrane topography of SARS-CoV-2-NSP4. The insets show the helical N-terminal signalling peptide (1-28 residues), helical protrusion of residues 305-300, and helical protrusion of the residues 367-371.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

Apart from the alpha defensin–like conserved domain, SARS-CoV-2-NSP4 was also predicted to share 41% domain identity with a fold recognized in the c-terminal domain of feline coronavirus from residues 403-494. Furthermore, the protein also shares 61% similarity with a fold recognized in the RNA-directed RNA polymerase enzyme of mouse hepatitis virus from the residues 410-499. Clinically relevant similarity was observed, although very low, with a protein of the human papilloma virus called the major capsid protein I1. Although a large extent of the SARS-CoV-2-NSP4 sequence was predicted not to match with presently known conformations, the protein was nevertheless predicted to fold into its native conformation attaining 49% alpha helix, 21% beta strand and 6% disordered regions. SARS-CoV-2-NSP4 was also predicted to contain a transmembrane topography contining constituting 25% TM helix with N-terminal signal peptide and C-terminal embedded in.

The transmembrane topography of the SARS-CoV-2-NSP4 revealed the possibility of containing five subunits (S1-S5), with a helical N-terminal signaling peptide stretching from residues 1-28. Figure 4B displays the helical nature of the protrusions of SARS-CoV-2-NSP4 from the viral body.

3.5 SARS-CoV-2-NSP5 (C-like proteinase)

The protein is most likely to share a folding pattern similar to the SARS-CoV-3CLPRO, which is functionally a trypsin-like cysteine protease, and serves as a template for 2-301 residues (Shown in Figure 5A: Rank1 with fold library id- *d2duca1*). The SARS-CoV-2-NSP5 also shares 50% identity (1-306 residues) to the human coronaviruses hku4 in complex with Michael acceptor sg85, and 45% with hcoV-NL63 3C-like protease (1-305 residues) of the organism transmissible gastroenterovirus. Furthermore, the NSP5 protein is likely to share domain identity from the residues 198-298 with the C-terminal of SARS-CoV main protease dimerized due to domain-swapping. Interestingly, the SARS-CoV-2-NSP5 also shares domain similarity (from residues 17-60) with poliovirus protease, 3CD protein (Shown in Figure 5B), which serves as the precursor to the RNA-dependent RNA polymerase enzyme. This is indicative that evolutionary pathways of the coronavirus and poliovirus ancestors overlapped in the past (Baric et al., 1990), and a thorough investigation might open new avenues to correlate the organisms to develop a vaccine . Intriguingly, the SARS-CoV-2-NSP shares a 62% identity (residues 11-42) with a domain of the immunosuppressant, human complement C5 complexed with tick inhibitors omci, r raci1 and cirpt1 (shown in Figure 5C).

Yet again it raises the question about what could be the link between the proteins involved human immunity and the SARS-CoV-2 non-structural proteins. As the protein most likely shares fold similarities with proteins of curable viral diseases like polio, bronchitis, etc, its domain analysis might unfold the insights for possible drugs and rapid-response vaccines.

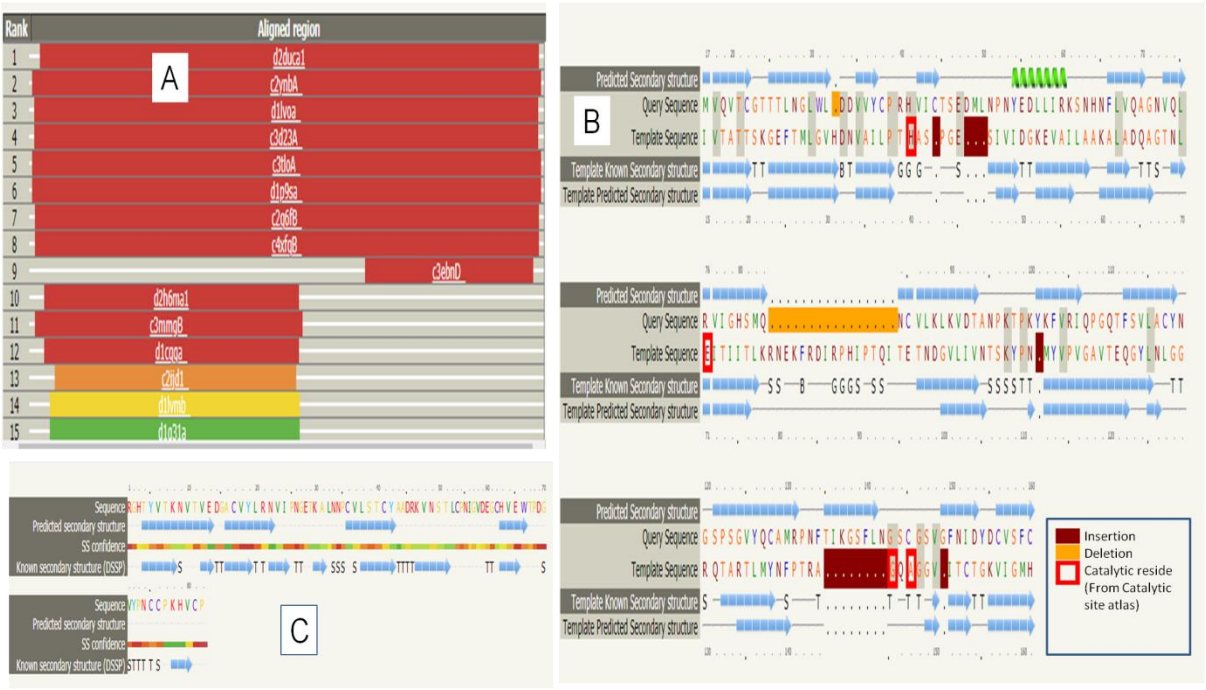


Figure 5. (A) Domain-fold analysis of SARS-CoV-2-NSP5 (C-like protease) showing a conserved domains; (B) Comparative structural analysis of SARS-CoV-2-NSP5 with poliovirus precursor protein, 3CD (PDB: 2IJD) showing the catalytic residues, insertion and deletion of amino acid sequences ; (C) Comparative structural analysis of SARS-CoV-2-NSP5 with human immonosupressant protein, Complement C5 complexed with tick inhibitors OmCI, RaCI1 and CirpT1 (PDB: 6RQJ).

Now, Returning to the hypothesis of evolution with number of infections, what seems intriguing is the chimeric nature of the SARS-CoV-2 proteins which are mostly distinct, yet contains conserved folds from a large pool of proteins from different species. Tracing the conserved domains might also reveal the nature of the infections and evolution thereby, of the corona viruses. It is fascinating to envisage that by modelling such evolutionary maps with historical investigation of epidemics, the future outbreak and the nature of the evolution of the virus could be predicted (Zhao et al., 2020).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

3.6 SARS-CoV-2-NSP6

Contrary to the former, the SARS-CoV-2-NSP6 appears to have evolved into a largely divergent protein and does not share identity with any domains of known protein conformations from the residues 1-225 (Shown in Figure 6A). However, the C-terminal of the protein loosely shares domain-folds with a variety of proteins from unrelated species like human cap-specific adenosine methyltransferase bound to SAH (Shown in Figure 6C), Capsid assembly protein VP3 of Birnaviruses, calcium-regulated actin bundling protein of *Dictyostelium discoideum*, etc. Furthermore, the protein appears to have an evolutionary fingerprint (from residues 236-261) shared with the photosynthetic RuBisco of the psychrophile, *Thalassiosira Antarctica* (shown in Figure 6B), which can survive in temperatures below the freezing point of sea water (Aletsee and Jahnke, 1992). The presence of extremely divergent evolutionary fingerprints of, for example, hyperthermophiles and psychrophiles, and the taxonomical variations of the organisms with which SARS-CoV-2 shares conserved domains, and the distinctly novel profile of some SARS-CoV-2 proteins, further puts our hypothesis of evolution with infections as a prime candidate for scrutiny. Indeed it is a possibility that the SARS-CoV-2 has evolved through a route consisting of large-scale infections, formation of a quasispecies pool, and the natural selection of the most advantageous form.

Furthermore, the topography of SARS-CoV-2-NSP6 was predicted to be a transmembrane protein with seven subunits (S1-S7) with the N-terminal (residues 1-21) involved in signalling and embedded C-terminal. The overall structure of the protein is predicted to contain 84% alpha helix, a small fraction of beta strand around ~1%, disordered region around 5% and TM helix (58%). The predicted transmembrane topography of SARS-CoV-2-NSP6 revealed that the presence of seven subunits (S1-S7) with the N-terminal (residues 1-21) as the signal peptide (Shown in Figure 6D). The conformation of the N-terminal is predicted to be alpha helical in nature.

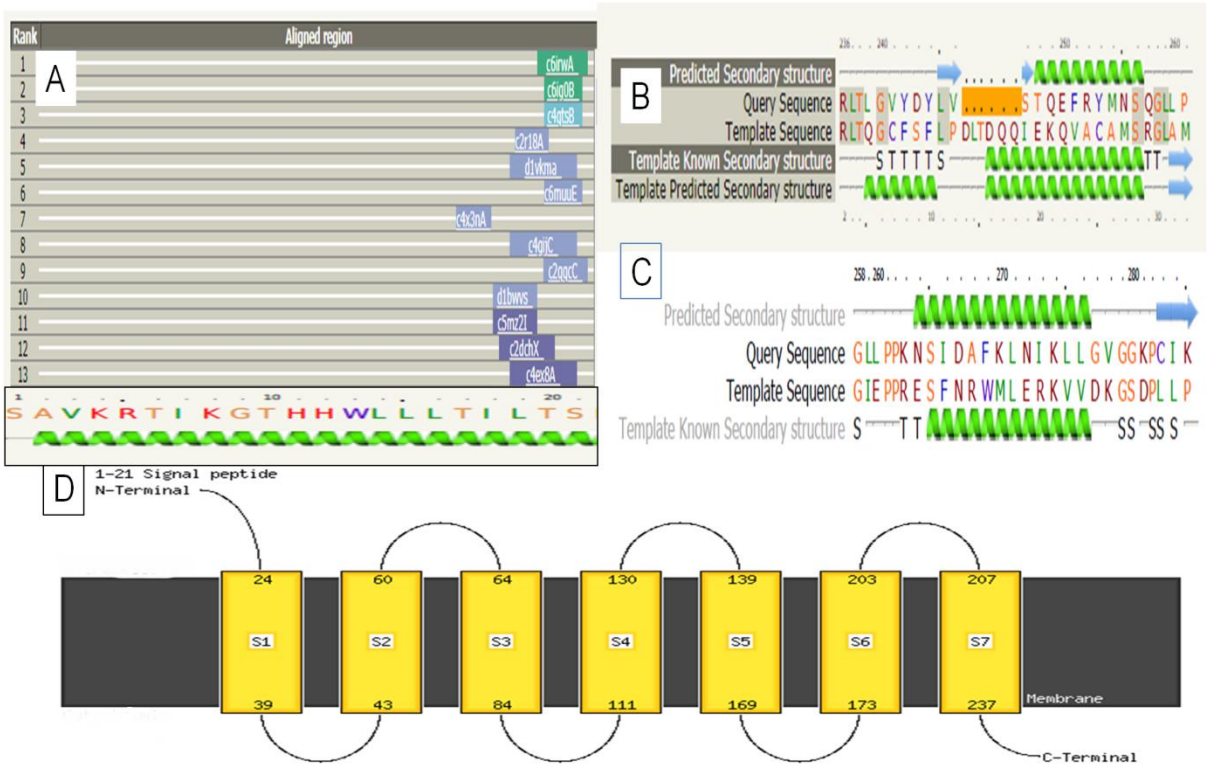


Figure 6. (A) Comparative domain-fold analysis of SARS-CoV-2-NSP6 showing conserved domain at the C-terminal whereas from residues 1-225 do not share identity with any protein structures; (B) Comparative structural analysis of SARS-CoV-2-NSP6 with photosynthetic RuBisco from *Thalassiosira Antarctica* (PDB: 5MZ2) (C) Comparative structural analysis of SARS-CoV-2-NSP6 with human cap-specific adenosine methyltransferase bound to SAH (PDB: 6IRW); (D) Transmembrane Topography of SARS-CoV-2-NSP6 showing seven subunits with N-terminal signalling peptide (1-21 residues). Inset shows the alpha helical structural predition of the signal peptide.

3.7 SARS-CoV-2-NSP7

It was found that the protein shares 99% identity with SARS-CoV-NSP7 with a single residue difference at position 70 (K->R) within residues 1-83 (Shown in Figure 7A). This is indicative of the nature of the point mutations occurring in genome of the SARS-CoV2 (Lysine- AAA/AAG; Arginine- AGA/AGG) which is a purine-purine transition. Presumably, one among many causes for the mutation could be recombination inside the host cell. With mutations occurring at fast rate, small variations introduced consequently might cumulatively result in the behavioural change of SARS-CoV-2 protein to modify or enhance protein function (Nieba et al., 1997). Figure 7B shows that small variations accumulate to a divergence 100% to 24%, which indeed is alarming. If we consider the hypothesis that every

infection would bring slight changes in the SARS-CoV-2 genome, gradually it will lead to the creation of several quasispecies (Domingo, 2002), and the divergence between may accumulate to form divergent proteins with enhanced or modified function.

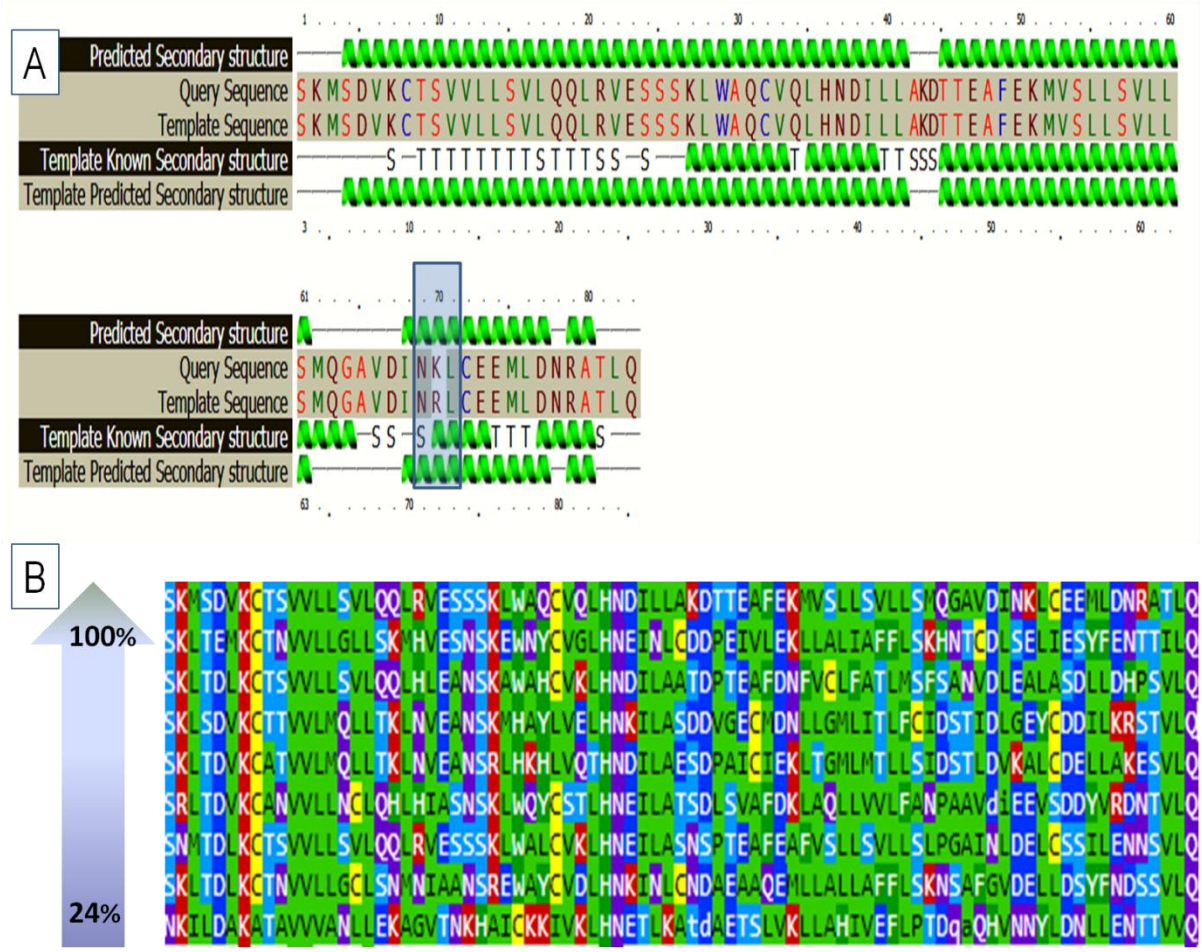


Figure 7. (A) Sequence alignment (residue 1-83) between SARS-CoV-NSP7 and SARS-CoV-2-NSP7 showing a single residue change at (K->R)(PDB: 1YSY); (B) Effect of point mutation on a sequence leading to a divergence from 100% to 24%.

Furthermore, the proteins shares domain identity with CRISPR-associated CasB of *Thermobifida fusca* (from residues 37-76) and exonuclease of *Methylocaldum szegediense* (residues 3-38), which are both known to be thermophillic in nature (Lammerts van Bueren et al., 2012). Symptomatically, COVID-19 relevant domain-identity was observed to be the evolutionary fingerprint of the superfamily, 'fumarate reductase respiratory complex transmembrane subunits' (Residues 52-81). In relation to immune response, the protein shares domain-fold identity with human C4b-binding protein (residues 34-52) associated with the immune system.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

3.8 SARS-CoV-2- NSP8

The SARS-CoV-2-NSP8 has been preedicted to be similar to the SARS-CoV NSP7-NSP8 hexadecamer with 97% similarity in fold recognition. Therefore the predicted three dimensional structure of SARS-CoV-2-nsp8 templated on the latter protein of SARS-CoV has been projected with 100% confidence. The sequence alignment with corresponding domain conformation revealed that most likely NSP8 of SARS-CoV-2 ORF1ab contains the conservation of largest number domains than other proteins of the open reading frame. Figure 8 highlights some of the conserved structural features between SARS-CoV-2-NSP8 and SARS-CoV-(NSP7-NSP8) hexadecamer.

The protein also shares 97 % fold similarity with corona virus NSP-8 superfamily, and also, 41% similarity to non-structural protein, nsp6, of feline coronavirus. On similar lines to the NSP2, NSP4 etc, the SARS-CoV-2-nsp8 is also predicted to share similar domains to other proteins across species, especially with several members of the human proteome including alpha helical domains of human fkbp5 protein, human symplekin, human methionine aminopeptidase, human dual specificity protein kinase mps1, angiogenesis inhibitor etc. It is alarming, although the percentage identity is extremely low, but it is indicative of conserved domains between the proteomes of SARS-CoV-2 and human. This study intends to lay forward the conjecture that SARS-CoV-2 is randomly searching conserved domains of the human proteome to bolster strong adaptation to the host physiology (Fang et al., 2005). It is verily emphasized that the SARS-CoV-2 is strongly put under thorough investigation on the lines of conserved conformational domains to realize the ways SARS-CoV-2 would utilize proteins of the host cell.

Structurally, SARS-CoV-2-NSP8 is most likely to attain 59% alpha helix, 16% beta strand, and 16% disordered region, and shows strong resemblance to the secondary structure of the template protein. Significant and drastic change due to point mutation was not predicted, however, it is not without possibility that crucial mutations might change the functionality of the proteins turning into more human-like. In this context, the hypothesis that a large portion of the non-coding human genome has been thought to be remnant of viral genomes integrating into human genome (Brechot et al., 1980; Rohan, 2017; Zapatka et al., 2020), it must be investigated whether there is any possibility of SARS-CoV-2 to change life cycle.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

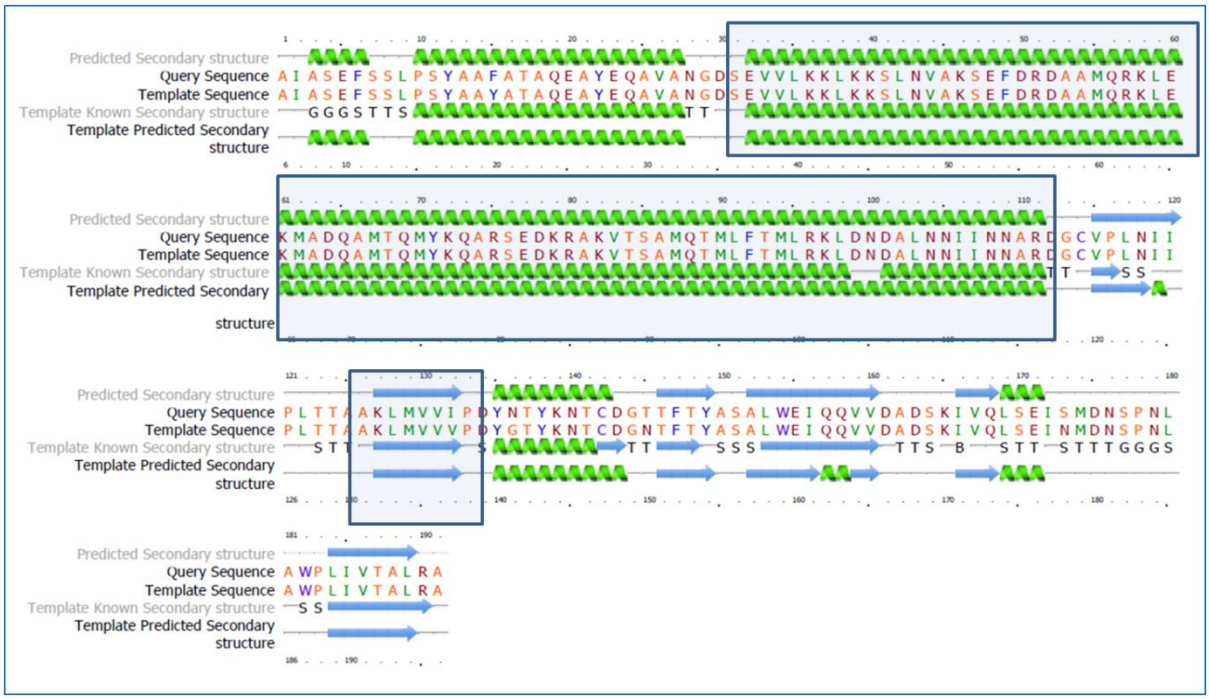


Figure 8. Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP8 and SARS-CoV NPS7-NSP8 hexadecamer (PDB:2AHM)

3.9 SARS-CoV-2-NSP9

SARS-CoV-2-NSP9 protein has been predicted to resemble SARS-CoV-NSP9 very strongly with 97% identity in fold recognition, with 100% coverage from residues 1-113 with predominantly beta sheet rich structure. The protein does not appear to exhibit drastic divergence from the Replicase NSP9 superfamily and possibly shares 97% identity with the Replicase NSP9 superfamily. The divergence has been predicted to be significant with RNA-binding human coronavirus-Replicase with 42% identity, and also 30% identity with porcine delta coronavirus NSP9. The domain-fold analysis revealed the presence of atleast a single conserved domain shared with multiple species (Show in Figure 9A), Interestingly, the protein also has the propensity to share identity, although very low, to a domain of Japanese encephalitis virus -E protein (Shown in Figure 9B). Therefore, the possibility of therapeutic molecules against e protein, whether approved or under investigation, to be also active to block SARS-CoV-2 replicase NSP cannot be ignored (Prompetchara et al., 2020). Furthermore, the protein also shares identity with the recognized folds of the coat protein of potato leafroll virus (51-74 residues) (Shown in Figure 9C) and a small hexapeptide stretch of der f23 allergen protein associated with dust allergy (Shown in Figure 9D).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

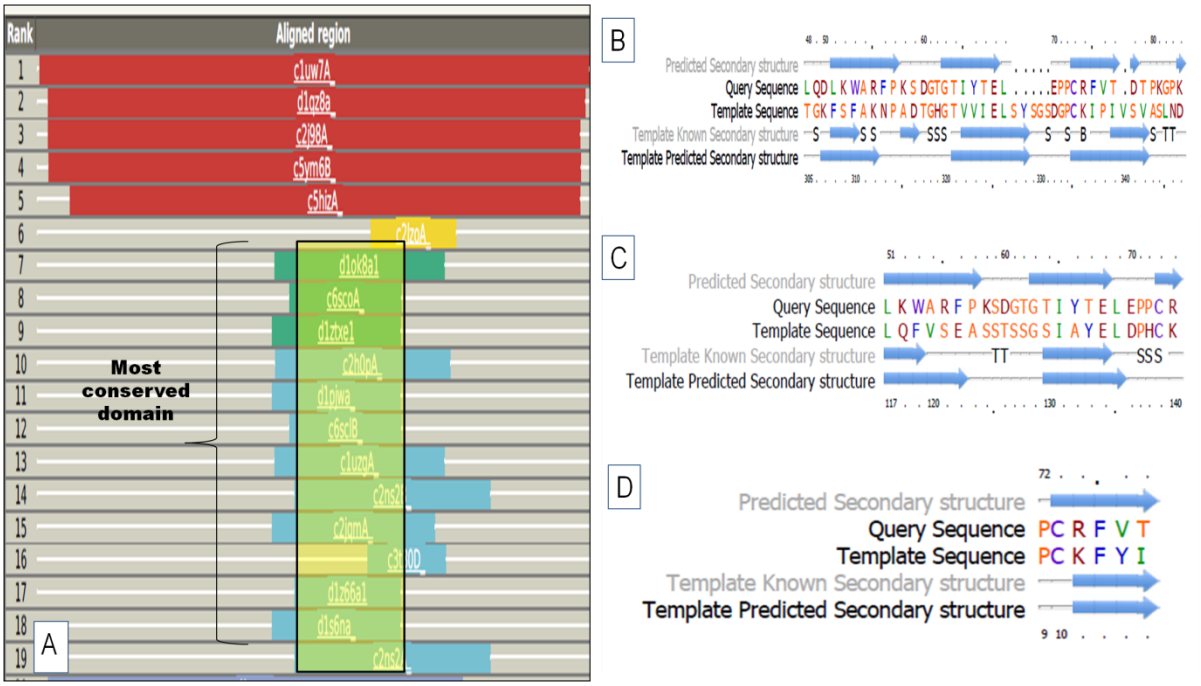


Figure 9. (A) Domain-fold analysis of SARS-CoV-2-NSP9 showing a conserved domain; **(B)** Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP9 and E protein of Japanese encephalitis virus (PDB: 5WSN); (C) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP9 and Coat protein (PDB:6SCO) of Potato leafroll virus (Scotland/strain 1/1984); (D) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP9 and dust mite allergen der f23 from *Dermatophagoides farinae* (Family: Hypothetical protein APE0525, N-terminal Domain).

3.10 SARS-CoV-2-NSP10

The structure of the SARS-CoV-2 NSP10 most likely attains functional native form resembling the folds specific to the superfamily of coronavirus NSP10-like proteins with 19% alpha helix, 32% beta strands, and 19% disordered regions. The possibility of similarity has been predicted to be 95% in the amino acid stretch 6-128 residues. It also shares similarity in fold recognition to MERS-CoV- NSP6/NSP10 complex (61% identity) in the residue stretch 11-130, and SARS-CoV protein NSP10 Chain T from the residues 9-129 (98% identity). The domain-fold analysis revealed the scattered presence of multiple conserved domains shared across several species (Shown in Figure 10A). Interestingly the protein also contains the possibility of containing integrated evolutionary fingerprint shared with two plant proteins, (i) epidermal patterning factor-like protein-9 from the residues 90-108 (Shown in Figure 110B)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

(ii) plant protein, PSI-LHCI, with the alignment from residues 27-77 (Shown in Figure 10C). This could be considered as an indication of coronavirus infection among plants in the past and a thorough investigation might as well unfold important information towards developing plant based medicines (Xu and Asakawa, 2020). What is alarming is that the protein also contains imprints of domain which are present in DNA-binding protein, Carbohydrate-binding protein, several antimicrobial proteins and also toxins.

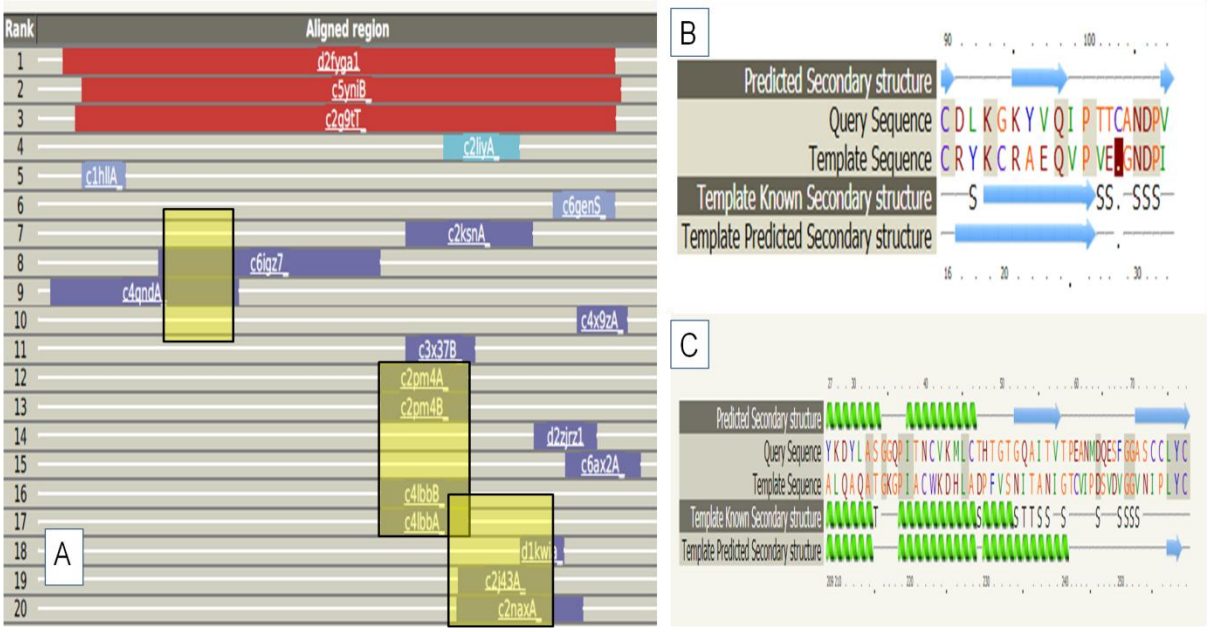


Figure 10. (A) Domain-fold analysis of SARS-CoV-2-NSP10 showing several scattered conserved domains; (B) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP10 and the plant peptide hormone, epidermal patterning factor-like protein-9 (of *Arabidopsis thaliana*) from the residues 90-108 (PDB: 2LIY); (C) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP10 and plant protein, PSI-LHCI (Organisms: *Bryopsis corticulans*, *Bryopsis plumosa*), with the alignment from residues 27-77 (PDB: 6IGZ).

3.11 SARS-CoV-2-NSP12 [RNA-dependent RNA Polymerase (RdRp)]

The predicted structure of the SARS-CoV-2-NSP12, i.e., RNA dependent-RNA polymerase (RdRp) of SARS-CoV-2 revealed that it is most likely to share 97% identity in domain-fold with SARS-CoV-NSP12 bound to NSP8 co-factor from the residues 118-918. It is most probable that the SARS-CoV-2-NSP12 would undergo a folding cascade similar to SARS-CoV-NSP12 to attain its native conformation. However, sequence alignment and structural prediction according to fold recognition revealed single amino acid changes across the two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

sequences, for instance, alanine in the SARS-CoV-NSP12 has been replaced by threonine at the 252 and 259 positions. Referring to the genetic code, Alanine is coded by ACU/ACC/ACA/ACG and threonine is coded by GCU/GCC/GCA/GCG, which implies a transition point-mutation of adenine to guanine. However, another mutation at the position 265, revealed a change from tyrosine (UAU/UAC) to lysine (AAA/AAG) which is purine-pyrimidine transversion. Moreover, the reported phylogenetic mapping of the SARS-CoV and SARS-CoV-2 genome revealed that they are divergent from a common ancestor (Tang et al., 2020). It is indicative that gradual accumulation of the point mutations in the common ancestor reservoir has concluded in the divergence, and the short timeline between the SARS-CoV epidemic (2002-2004) and SARS-CoV-2 pandemic (2019-2020) is alarming. The domain-analysis revealed the presence of a predominant conserved domain (shown in Figure 11A). A notable similarity of a domain of SARS-CoV-2-RdRp and SARS-CoV-RdRp is shown in Figure 11B.

The predictions about the secondary structures of the RNA-Dependent-RNA-Polymerase shows 3_{10} helix, hydrogen bonded turns, residues in isolated β-bridge, and bends. The native conformation of the protein is predicted to have 55% alpha helices, 7% beta sheets and 11% disordered regions which have been predicted by templating with SARS-CoV-RdRp. Although primarily the structural change in between the predicted secondary structure of the SARS-CoV-2 from SARS-CoV's experimentally verified structure appears to be mostly a frameshift. However, certain residue changes indeed are capable of engendering significant changes in secondary structure, for instance, the residue alterations at the positions 841 and 842 (LM in SARS-CoV to GC in SARS-CoV-2) increases the propensity to diverge from helical to a beta sheet structure (Shown in Figure 11C). This again indicates the potential of the point mutations for causing drastic and novel divergence in the folding patterns of proteins whose origins could be traced back to a common ancestor not so long ago. This possibility of the structural changes in the fold might result in the increment/ decrement or change of functionalities as they diverge evolutionarily.

Apart from sharing conformational features with SARS-CoV-NSP12 spanning 80-residues, the SARS-CoV-2-NSP12 also was predicted to contain a conserved domain spanning residues 456-903. The domain appears to be shared by other proteins including human rhinovirus' RdRp, EV71 RdRp complexed with GTP (human enterovirus), 3dpol RdRp (Sicinivirus), porcine aichi virus polymerase, murine norovirus' RdRp, Precursor protein

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

3CD of poliovirus etc. The predicted structural feature of the conserved domain is shown in figure.

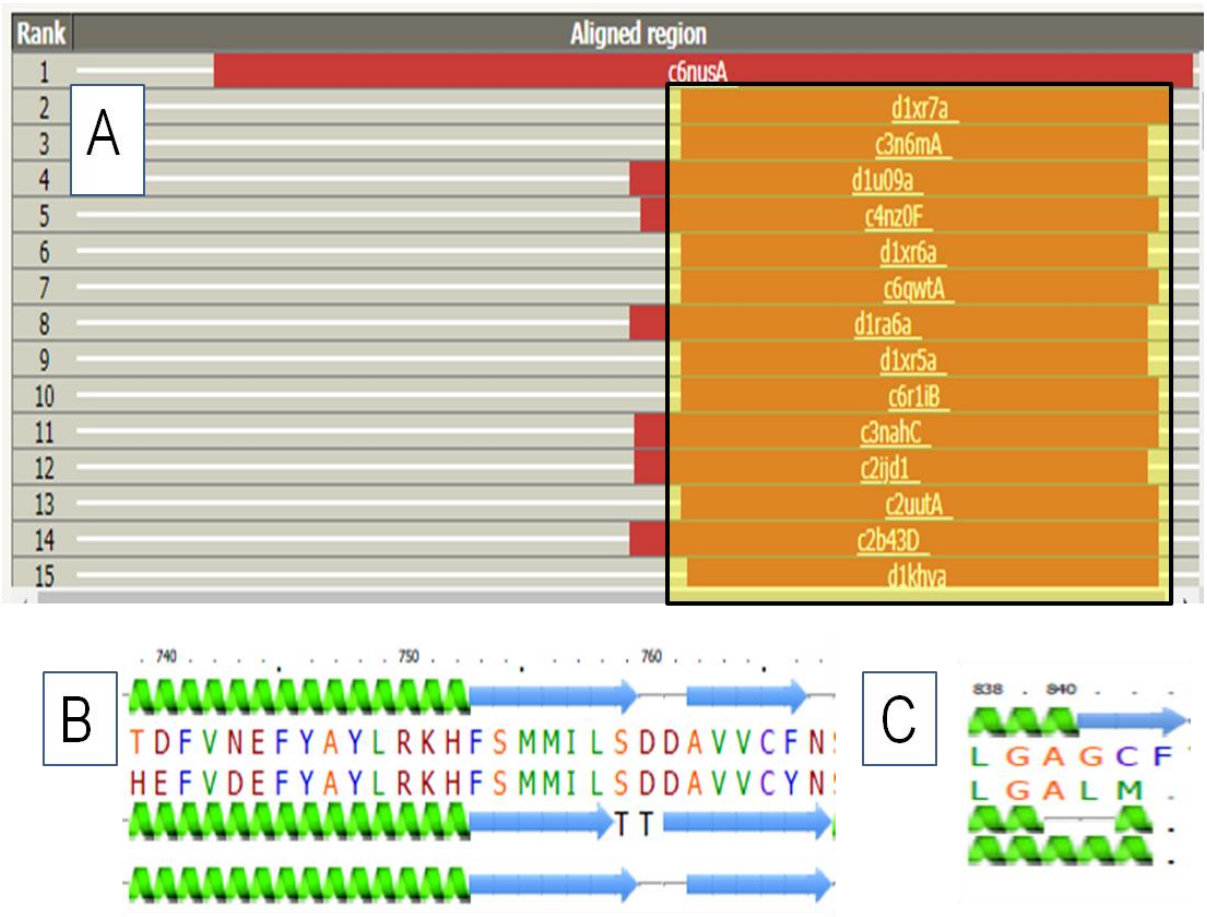


Figure 11. (A) Domain-fold analysis of SARS-CoV-2-NSP12 (RdRp) showing the presence of a predominant conserved domain; (B) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP12 and SARS-CoV-RdRp (PDB:6NUS); (C) Comparative structural analysis of the conserved domain between SARS-CoV-2-NSP12 and SARS-CoV-RdRp showing the increased propensity of alpha helix/beta strand transition with amino acid sequence change (PDB:6NUS).

3.12 SARS-CoV-2 -NSP13 (Helicase)

The three dimensional folded structure of the SARS-CoV-2 -NSP13 helicase has been predicted by templating against MERS-CoV helicase (MERS-CoV-NSP13). The BLASTP results indicated that the SARS-CoV-2 helicase is 72% similar to MERS-CoV-NSP13 and the three dimensional structure has been modelled with 100% confidence. The protein is also similar to another variant of the MERS-CoV-NSP13 with which it shares 73% identity. The protein also shares similar domains with significantly distant proteins like RNA binding

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

proteins and hydrolases of distantly related organisms. Figure 12 highlights the presence of a predominant conserved domain shared across several species which are distantly related exhibiting domain-fold similarity as low as 7% to 36 %.

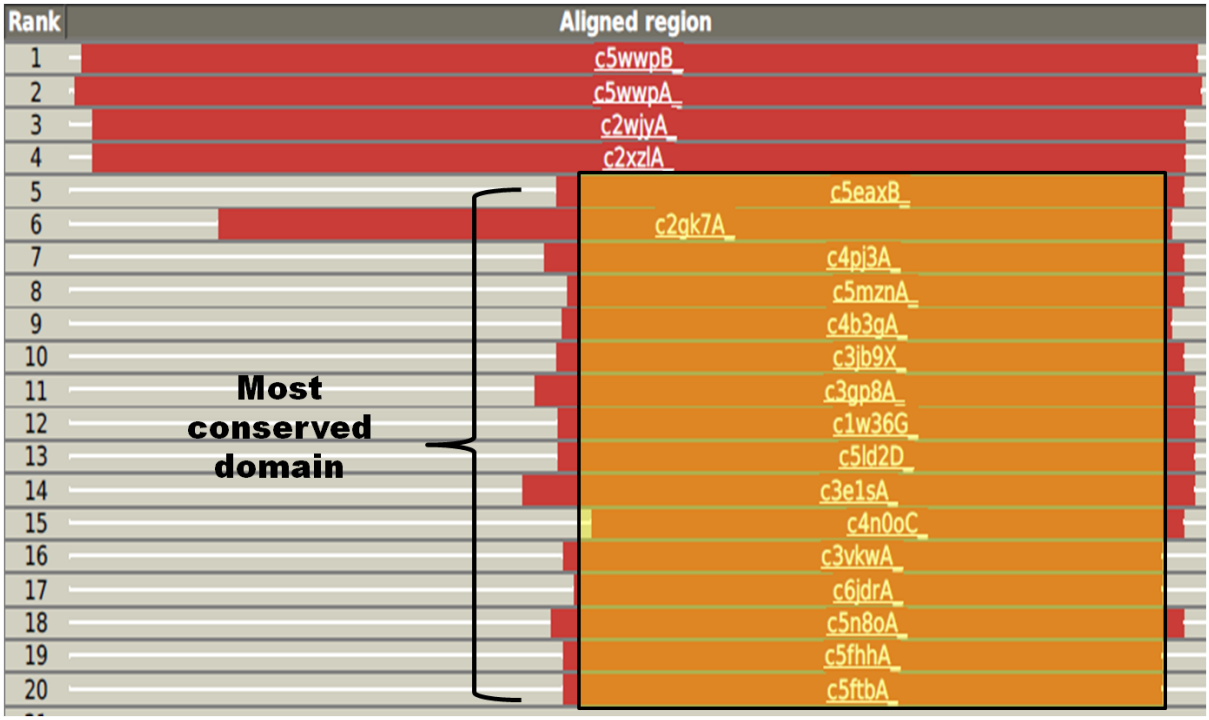


Figure 12. Domain-fold analysis of SARS-CoV-2-NSP13 (Helicase) showing the presence of a predominant conserved domain.

This is indicative that the molecular evolution of the protein has been taking place with other proteins of several species with conserved domains yet diverged significantly from common ancestors. The variation between predicted secondary structures of the SARS-CoV-2 helicase with the template (MERS-CoV-NSP13) revealed the possibility of the introduction of small structural variations through point mutations. For instance, the amino acid residue change in the position 41 (L to M) and 46 (N to S) increases the possibility of introducing structural change in the protein. It is well known that the proteins over time evolve gradually with small variations introduced in the amino acid sequence. This gradual process does not occur to be significantly alarming, but, considering the hypothesis held in this study to be true, it can be speculated that with the introduction of point mutations along with increasing number of infected individuals, the functionality of the protein might steeply enhance with greater affinity towards the substrate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

Similar to SARS-CoV-2-NSP12, NSP13 also appears to exhibit a conserved domain spanning the residues 240-260. The conserved domain shares similarities with human UPF1 helicase core, pre-mRNA-splicing factor CWF11 (Yeast spliceosome), arterovirus NSP10, Large subunit (helicase) of tomato mosaic virus, reverse gyrase from the *Thermotoga maritima* etc. Intriguingly, domain also shares an evolutionary fingerprint with RecD2 (Organism: *Deinococcus radiodurans*) which is associated with unwinding of DNA (Shadrick and Julin, 2010).

3.13 SARS-CoV-NSP14 (Exonuclease)

The fold recognition analysis of SARS-CoV-2-NSP14 revealed that the protein identifies most strongly with SARS-CoV's guanine-n7 methyltransferase with similar domain-folds spanning residues 1-525. According to the three dimensional folds predicted, the protein possibly exhibits 38% alpha helix, 19% beta strand, and 6% disordered region, with features like 3_{10} helix, hydrogen bonded turns, bends and loops. The domain analysis of the SARS-CoV-2-NSP14 revealed the presence of scattered conserved domains shared with several distantly related proteins (Shown in Figure 14).

The SARS-CoV-2-NSP14 appears to contain a conserved domain from the residues 2-59 shared with conformational variants of penton base protein of adenoviruses (Shown in Figure 14A). Residues 197-236 exhibit a domain fold that shares conserved imprint with an apoptosis-associated protein family called the inhibitor of apoptosis (IAP) repeat. A Snake toxin-like domain fold appears to identify with residue stretch from 411-427 from the SARS-CoV-2-NSP14 with 47% similarity. Furthermore, residues from 48-160 shares identity, interestingly, with evolutionarily unrelated proteins like the avian reovirus inner capsid protein sigma a (Shown in Figure 14B), chain A of poly [ADP-ribose] polymerase rcd1 from the plant *A.thaliana*, and with 50S ribosomal protein of *Pseudomonas aeroginosa*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

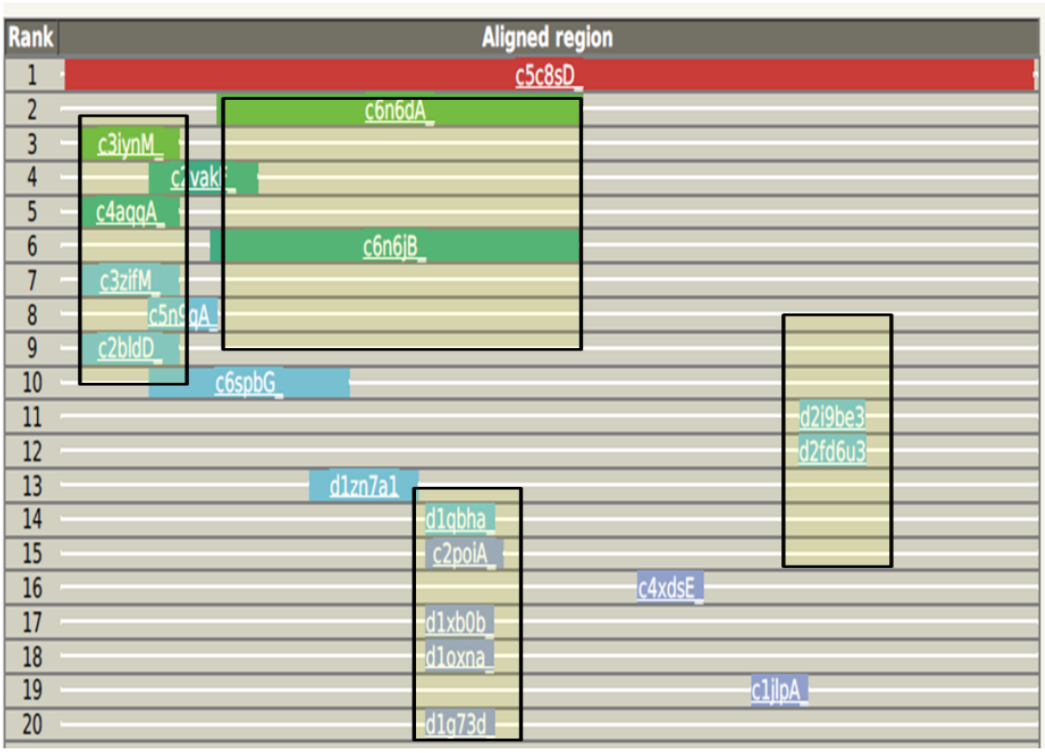


Figure 13. Domain-fold analysis of SARS-CoV-2-NSP14 (Exonuclease) showing the presence of scattered conserved domains.

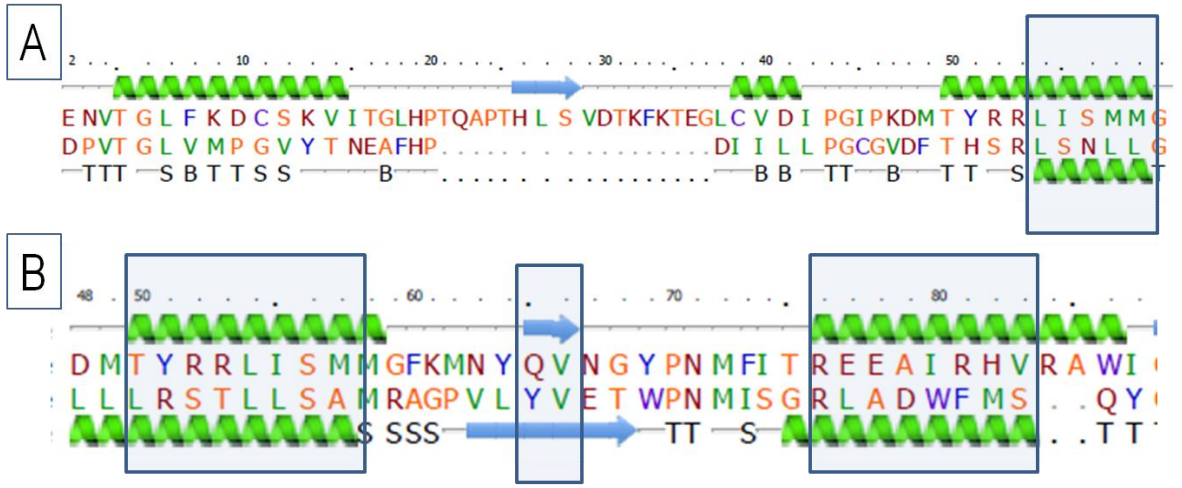


Figure 14. (A) Conserved conformation of domains between SARS-CoV-2-NSP14 and penton base protein of human adenovirus type 5 (PDB: 3IYN- replaced by 6B1T)); (B) Conserved conformation of domains between SARS-CoV-2-NSP14 and Avian reovirus inner capsid protein Sigma a (PDB: 2VAK).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

3.14 SARS-CoV-2-NSP15 (Endoribonuclease)

The SARS-CoV-2-NSP15/Endoribonuclease also was observed to contain scattered conserved domains indicating evolutionary relationship with diverse organisms (Shown in Figure 15A). Nevertheless, SARS-CoV-2-NSSP15 has been predicted to share maximum conformational similarity (88%) with the uridylate-specific endoribonuclease along with domain specific similarity (51%) with MERS-CoV-NSP15 which is a hydrolase in function. The structural predictions reveals 16% alpha helix, 48% beta strands and 18% disordered region. The alignment of the amino acid sequences of the two aforementioned proteins also reveals the possibility of point mutations leading to change in folding pattern. For example, the shift of MERS-CoV's tyrosine at position 7 to phenylalanine in SARS-CoV-2 at the corresponding aligned region increases the propensity to form beta sheets which otherwise is known to fold into helical structure (Emruzi et al., 2018) (Shown in Figure 15B). This is again strongly speculative that the intrinsic amino acid sequence of the SARS-CoV has chosen an evolutionary path where a slight change is capable of triggering different folding pathways.

Furthermore, SARS-CoV-2-NSP15 appears to have domain-fold similarity to putative septation protein spovg from *Staphylococcus epidermidis*, human cytosolic aconitase2, and also with several viral proteins like outer protein of bacteriophage T4 isometric capsid . The alarming possibility is that SARS-CoV-2 proteins, even though share insignificant similarity to human proteome, might contain a reservoir of similar domains to the human proteins. Some of the similar domains are not without the possibility of mimicking the enzymatic site as well as substrate-like conformation/chemistry of human proteins. It is already established through various reports that the spike protein of the Coronavius family binds strongly with ACE2 receptors. It has also been found that the Spike protein of SARS-CoV-2 has a stronger binding to ACE2 receptor than its SARS-CoV parallel. This is indicative of the evolution of this viral family to undergo enhanced adaptation with the human genome and could be speculated as a case of co-evolution (Zheng and Perlman, 2018). At this point, summoning the hypothesis strongly held in this study it could be speculated that proteins of the SARS-CoV-2 will evolve more divergently with perhaps enhanced affinity towards human cell substrates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0286.v1

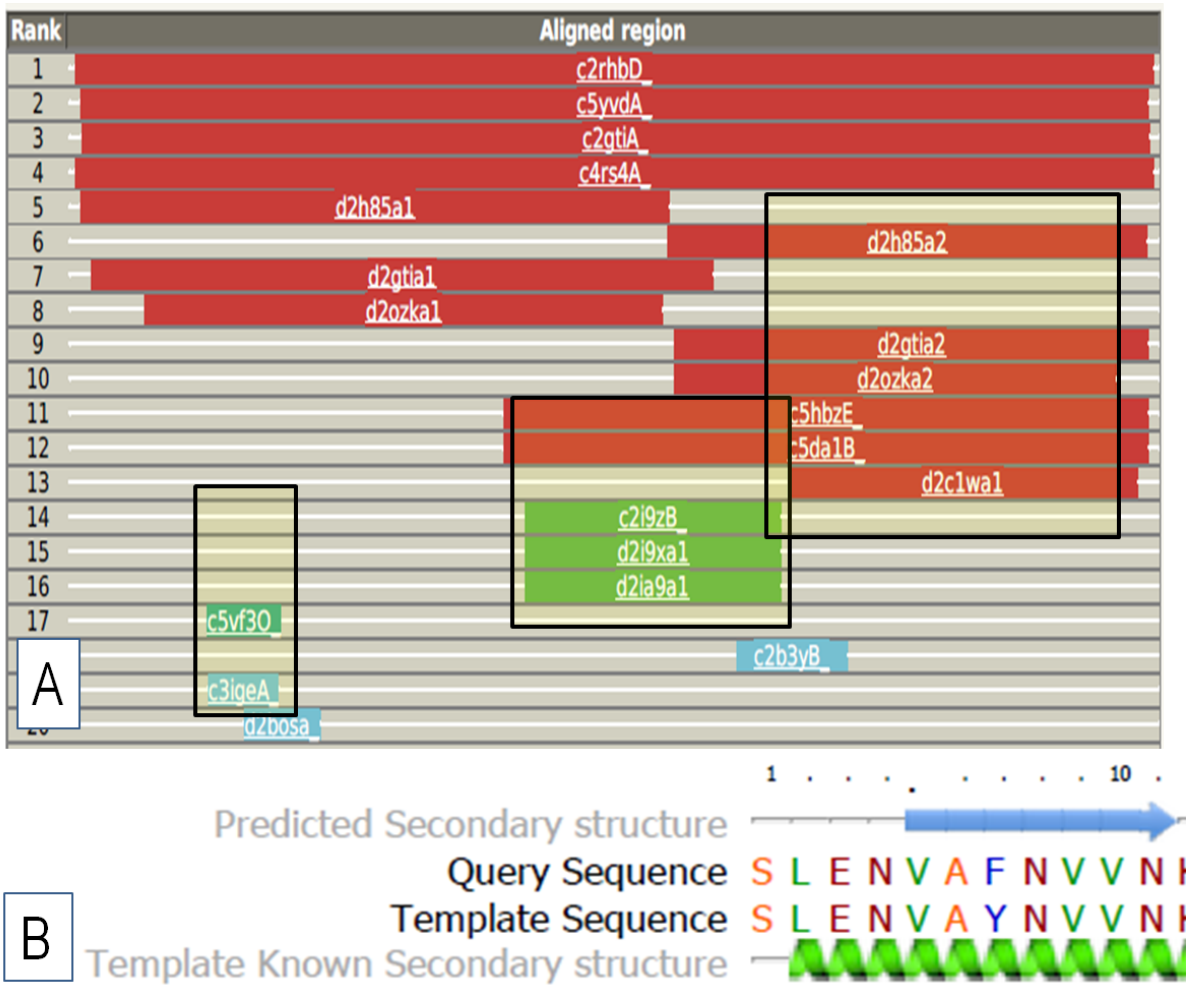


Figure 15: (A) Domain-fold analysis of SARS-CoV-2-NSP14 (Exonuclease) showing the presence of scattered conserved domains; (B) Excerpt from the comparative domain-fold analysis between SARS-CoV-2-NSP15 and Uridylate-specific endoribonuclease from NSP15-H234A mutant (Organism: SARS-CoV) displaying a single amino acid difference (F->Y) modulates the sequentially similar domains to fold differently (PDB: 2RHB).

3.15 SARS-CoV-2-NSP16 (O-Ribose-methyltransferase)

The folds of SARS-CoV-2-NSP16 have been modelled primarily with the template of SARS-CoV NSP10/NSP16 complex which appears to cover 97% residues from 3-293 (Shown in Figure 16A). Similarly the protein also shares 66% domain-specific similarity with MERS-CoV NSP10/NSP16 complex covering 1-294 residues. Interestingly, the protein most likely shares domain-specific similarity with protein of various other species, like the protozoan protein-ribosomal RNA methyltransferase (46-214 residues), yeast tRNA methyltransferase trm72/trm734 complex which is crucial for 2'-O-methylation at the fist position of anticodon

in specific tRNAs (50-216 residues), etc. Most alarming is its fold similarity with a protein of thermophilic archea, *Thermoplasma volcanicum* gss1 methyltransferase which is associated to cell division (50-214 residues) (Shown in Figure 16B). Although the identity shared is most likely 19%, yet it is particularly crucial to consider the possibility that the chimeric SARS-CoV-2 genome on its evolutionary pathway might also integrate useful mutations and adapt to extreme conditions for survival.

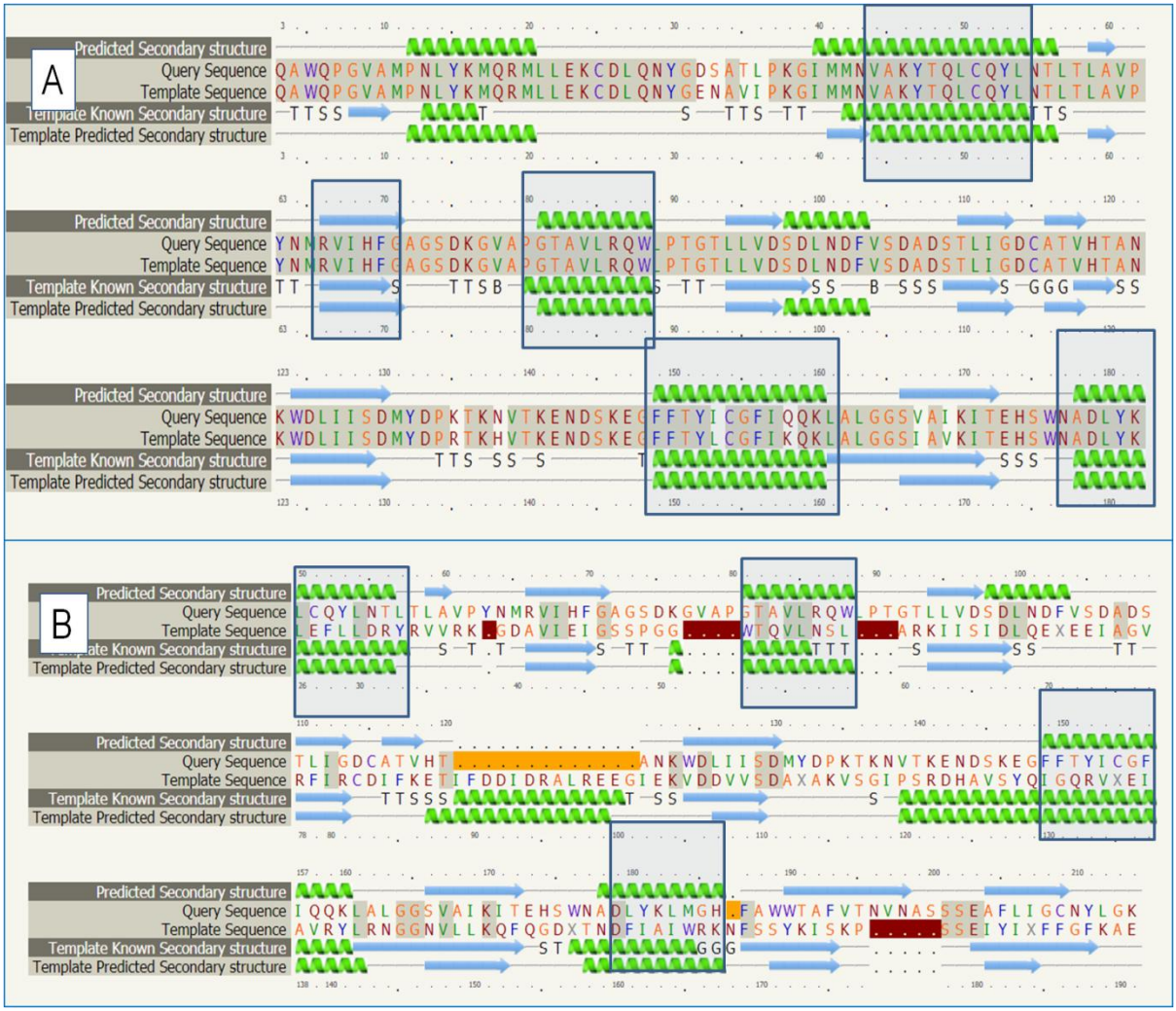


Figure16: (A) Conserved domain conformation between SARS-CoV-2-NSP16 and NSP10/NSP16 complex of SARS-CoV (PDB: 3R24); (B) Conserved domain conformation between SARS-CoV-2-NSP16 and Ribosomal RNA large subunit methyltransferase (PDB: 3DOU) from *Thermoplasma volcanicum* gss1.

The proteomic similarities of SARS-CoV-2 to a wide spectrum of proteins across several species ranging from human to other curable disease-associated pathogens, projects two possibilities: Firstly, the novel chimeric forms of some of the SARS-CoV-2 proteins is

indicative of rapid evolution, and the possibility of future emergence of a novel descendent cannot be ignored. Secondly, with the understanding that proteins conserve domains more strongly than the amino acid sequence, the analysis of conserved domains of the SARS-CoV-2 proteins might shed light on functional annotation and evolution, and the similarity of the SARS-CoV-2 protein domains with curable viral diseases could be exploited for deriving therapeutic insights.

4. CONCLUSION

The study investigated the folding patterns of SARS-CoV-2 polyprotein, ORF1ab, essential for the pathogenesis of the virus inside the host cell, by employing Phyre 2 fold recognition serve. The study specifically focuses on domain-wise fold recognition to generate evolutionary insights into the conformation and function of the SARS-CoV-2 ORF1ab. The rationale adopted was that the functionality of proteins depends largely on their native folded structure, which in turn, predominantly depends upon the intrinsic amino acid sequences. Therefore, accumulation of slight changes in the genome is expected to modify or enhance the functionality of the protein, by engendering small changes in the folding patterns. The results of the study indicated the presence of two classes of proteins encoded by ORF1ab of SARS-CoV-2 based on the structural variation. First class constitutes the proteins which seem to be conformational variants of polypeptides encoded by related organisms, like SARS-CoV and MERS-CoV; whereas, the second class appears to be highly chimeric and divergent which share conserved domain-folds with evolutionary distant proteins from other species, like psychro- and thermophillic organisms, various members of the plant kingdom, members of the phylogenetically close other corona viruses, and also curable disease-associated viruses like the bronchitis virus. The study holds the view very strongly that these domain-folds are evolutionary fingerprints which were integrated in the common ancestors of SARS-CoV-2 by infecting distantly related organisms in the past. It is hypothesized thererby that the ancestors of SARS-CoV-2 co-evolved with other species by causing wide-spread infection, integrating vital signals from the genome of the host cells, and gradually accumulating small variations to evolve with modified or enhanced protein functionalities. The study undertook an unconventional approach to keep a biased assumption by holding the hypothesis to be true, and attempted to anticipate the creation of a large pool of SARS-CoV-2 quasispecies with the exponential rise in the number of COVID-19 infected individuals. Accordingly, it is speculated that slight variations in domain folding patterns of SARS-CoV-2 proteins could be acquired with every new infected individual, giving rise to the creation of

multiple reservoirs of drastic and novel protein functions. Therefore, it is strongly recommended that the hypothesis be investigated experimentally and thoroughly modeled to accurately project the present solution and predict the possibility of future emergence. It is interesting to envision that a mathematical model would be developed to predict the nature of the future CoV taking into account (i) the conserved domain folding patterns, (ii) considering the past and present conformational forms of the proteins as intermediates, (iii) identifying the possible natural select, (iv) Mapping the enhanced protein functions attained through the process of evolution, (ii) Considering the regions with maximum number of SARS-CoV-2 infections as possible hotspots for future emergence. On positive lines, this study is driven to emphasize on thorough investigation of SARS-CoV-2 protein folding patterns to target enhanced protein functionalities. We also believe that The SARS-CoV-2 domains conserved in other species like poliovirus and bronchitis virus might shed light on designing strategies to develop vaccines and drugs.

Author Contributions

Srijeeb Karmakar: Conceptualization, Methodology, Formal analysis. **Sachin Kumar:** Validation, Supervision. **Vimal Katiyar:** Supervision, Funding acquisition, Resources, Project administration

Declaration of interest: Authors declare no conflict of interest

Acknowledgement

The authors are grateful to **Mr. Satadru Chakraborty** for assisting in troubleshooting.

References

Aletsee, L., Jahnke, J., 1992. Growth and productivity of the psychrophilic marine diatoms Thalassiosira antarctica Comber and Nitzschia frigida Grunow in batch cultures at temperatures below the freezing point of sea water. Polar biology 11, 643-647.

Baric, R.S., Fu, K., Schaad, M.C., Stohlman, S.A., 1990. Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. Virology 177, 646-656.

Brechot, C., Pourcel, C., Louise, A., Rain, B., Tiollais, P., 1980. Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma. Nature 286, 533-535.

Canutescu, A.A., Dunbrack, R.L., 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein science 12, 963-972.

de Haan, C.A., Haijema, B.J., Masters, P.S., Rottier, P.J., 2008. Manipulation of the coronavirus genome using targeted RNA recombination with interspecies chimeric coronaviruses, SARS-and Other Coronaviruses. Springer, pp. 229-236.

Dobson, C.M., 2003. Protein folding and misfolding. Nature 426, 884.

Domingo, E., 2002. Quasispecies theory in virology. Journal of Virology 76, 463-465.

Emruzi, Z., Aminzadeh, S., Karkhane, A.A., Alikhajeh, J., Haghbeen, K., Gholami, D., 2018. Improving the thermostability of Serratia marcescens B4A chitinase via G191V site-directed mutagenesis. International journal of biological macromolecules 116, 64-70.

Fang, S.G., Shen, S., Tay, F.P., Liu, D., 2005. Selection of and recombination between minor variants lead to the adaptation of an avian coronavirus to primate cells. Biochemical and biophysical research communications 336, 417-423.

Huber, R., Langworthy, T.A., König, H., Thomm, M., Woese, C.R., Sleytr, U.B., Stetter, K.O., 1986. Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C. Archives of Microbiology 144, 324-333.

Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics 77, 499-508.

Jain, N., Shankar, U., Majee, P., Kumar, A., 2020. Scrutinizing the SARS-CoV-2 protein information for the designing an effective vaccine encompassing both the T-cell and B-cell epitopes. bioRxiv.

Jefferys, B.R., Kelley, L.A., Sternberg, M.J., 2010. Protein folding requires crowd control in a simulated cell. Journal of molecular biology 397, 1329-1338.

Karmakar, S., Sharma, L.G., Roy, A., Patel, A., Pandey, L.M., 2018. Neuronal SNARE complex: A protein folding system with intricate protein-protein interactions, and its common neuropathological hallmark, SNAP25. Neurochemistry international.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols 10, 845.

Lammerts van Bueren, A., Otani, S., Friis, E.P., Wilson, K.S., Davies, G.J., 2012. Three-dimensional structure of a thermophilic family GH11 xylanase from Thermobifida fusca. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 68, 141-144.

Li, X., Giorgi, E.E., Marichann, M.H., Foley, B., Xiao, C., Kong, X.-p., Chen, Y., Korber, B., Gao, F., 2020. Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. bioRxiv.

Liu, Y., Gayle, A.A., Wilder-Smith, A., Rocklöv, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine.

Ma, B.-G., Chen, L.-L., Zhang, H.-Y., 2007. What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms. Journal of molecular biology 370, 439-448.

Ménade, M., Kozlov, G., Trempe, J.-F., Pande, H., Shenker, S., Wickremasinghe, S., Li, X., Hojjat, H., Dicaire, M.-J., Brais, B., 2018. Structures of ubiquitin-like (Ubl) and Hsp90-like domains of sacsin provide insight into pathological mutations. Journal of Biological Chemistry 293, 12832-12842.

Neher, R.A., Dyrdak, R., Druelle, V., Hodcroft, E.B., Albert, J., 2020. Potential impact of seasonal forcing on a SARS-CoV-2 pandemic. Swiss Medical Weekly 150.

Nieba, L., Honegger, A., Krebber, C., Plückthun, A., 1997. Disrupting the hydrophobic patches at the antibody variable/constant domain interface: improved in vivo folding and physical characterization of an engineered scFv fragment. Protein engineering 10, 435-444.

Porcheddu, R., Serra, C., Kelvin, D., Kelvin, N., Rubino, S., 2020. Similarity in Case Fatality Rates (CFR) of COVID-19/SARS-COV-2 in Italy and China. The Journal of Infection in Developing Countries 14, 125-128.

Prompetchara, E., Ketloy, C., Palaga, T., 2020. Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. Asian Pac J Allergy Immunol 38, 1-9.

Qiu, Y., Xu, K., 2020. Functional studies of the coronavirus nonstructural proteins. STEMedicine 1, e39-e39.

Remmert, M., Biegert, A., Hauser, A., Söding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 9, 173.

Rohan, S., 2017. Ancient Viruses Found in DNA. Chicago Youth Science Journal, 33.

Rotkiewicz, P., Skolnick, J., 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of computational chemistry 29, 1460-1465.

Shadrick, W.R., Julin, D.A., 2010. Kinetics of DNA unwinding by the RecD2 helicase from Deinococcus radiodurans. Journal of Biological Chemistry 285, 17292-17300.

Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., 2020. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clinical Infectious Diseases.

Söding, J., 2004. Protein homology detection by HMM–HMM comparison. Bioinformatics 21, 951-960.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell.

Weber, A., Ianelli, F., Goncalves, S., 2020. Trend analysis of the COVID-19 pandemic in China and the rest of the world. arXiv preprint arXiv:2003.09032.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., 2020. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharmaceutica Sinica B.

Xu, Z., Asakawa, S., 2020. Can the novel coronavirus transmit via RNAs without protein capsids?

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q., 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444-1448.

Yang, C.-W., Chen, M.-F., 2020. Composition of human-specific slow codons and slow di-codons in SARS-CoV and 2019-nCoV are lower than other coronaviruses suggesting a faster protein synthesis rate of SARS-CoV and 2019-nCoV. Journal of Microbiology, Immunology and Infection.

Yi, H., 2020. 2019 novel coronavirus is undergoing active recombination. Clinical Infectious Diseases.

Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., Moch, H., Cooper, C.S., 2020. The landscape of viral associations in human cancers. Nature Genetics 52, 320-330.

Zhang, Y., Zhang, W., Ogata, S., Clements, D., Strauss, J.H., Baker, T.S., Kuhn, R.J., Rossmann, M.G., 2004. Conformational changes of the flavivirus E glycoprotein. Structure 12, 1607-1618.

Zhao, Z., Zhu, Y.-Z., Xu, J.-W., Hu, Q.-Q., Lei, Z., Rui, J., Liu, X., Wang, Y., Luo, L., Yu, S.-S., 2020. A mathematical model for estimating the age-specific transmissibility of a novel coronavirus. medRxiv.

 doi:10.20944/preprints202004.0286.v1

Zheng, J., Perlman, S., 2018. Immune responses in influenza A virus and human coronavirus infections: an ongoing battle between the virus and host. Current opinion in virology 28, 43-52.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., 2020. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine.