An *insilico* study to identify hidden features of Spike protein and Main protease of SARS-Cov2

Yogeshwar V Dhar[1,2], Priti Prasad[1,2], Nikita Tiwari[1], Vaishali Pankaj[1], Nasreen Bano[1,2], Sumit K Bag[1,2,*] , Mehar H Asif[1,2,*]

[1]CSIR-National Botanical Research Institute (CSIR-NBRI), Rana Pratap Marg, Lucknow-226001, INDIA

[2]Academy of Scientific and Innovative Research (AcSIR), CSIR-National Botanical Research Institute Campus, Rana Pratap Marg, Lucknow -226001, India

Running title: Structural analysis of SARS-Cov-2

***Correspondence:**
MHA (mh.asif@nbri.res.in)
SKB (sumit.bag@nbri.res.in)

CSIR-National Botanical Research Institute (CSIR-NBRI),
Rana Pratap Marg,
Lucknow-226001, INDIA
Tel: 91-522- 2297914

Abstract:

Motivation

The SARS-Cov-2 pandemic has gripped the entire world and a race to find either a cure or a vaccine for this pandemic is on. The public databases have a deluge of information in terms of genomic sequences and protein structures making it possible to study the minute details in terms of conserved motifs and super-motifs in its proteins

Results

In this study we have identified the hidden features of the Spike protein and the Main protease (Mpro) of SARS-Cov2. These domains have been identified using the standard bioinformatics tools and the pfam database. We found four domains reported in the pfam database are present in the spike protein and the Mpro of SARS-Cov2 but have not been reported earlier. These domains are specific to human strains of SARS-Cov2 and are not present in SARS-Cov or the coronaviruses of other animals. Using RIN we also identified the motifs and super-motifs in these two proteins that are important in understanding species wise changes as well as evolution driven variation in amino acids. Our results highlight several interesting features of the spike protein and Mpro of SARS-Cov2 that can be exploited for the development of various drug and vaccine therapies.

Contact: mh.asif@nbri.res.in, sumit.bag@nbri.res.in

Supplementary information:

Keywords: SARS-Cov-2, main protease, Spike protein, RIN

Introduction

The Covid-19 pandemic is on a roller coaster ride and has till date infected close to 7 million people worldwide leaving several thousands dead. The various measures taken to slow down the infection rate worldwide has resulted in major social as well as economic losses. There is an urgent need to develop therapeutic drugs as well as vaccines to fight this virus. Since the outbreak of this coronavirus, a number of studies are going on and published, focusing on the different aspects of this deadly virus including the mechanism of virus infection, treatment of symptoms, transmission, diagnosis, prevention and other clinical case studies in more than 20,000 reports (source: LitCovid, https://www.ncbi.nlm.nih.gov/research/coronavirus/).To combat the virus it is important to study its evolutionary patterns and the changes resulting in the proteins it encodes.

Coronaviruses, member of *coronaviridae* family, are a group of multifarious positive strand RNA viruses. The basic structure of this virus is consists of 4 structural proteins, spike(S), envelope(E), membrane (M) and nucleocapsids (N). Apart from these main structural proteins the virus also encodes for other accessory proteins of which chymotrypsin-like protease (3 CL protease) or the main protease in important for the processing of various polypeptides. Of the four structural proteins the M and E are involved virus assembly and the S protein is involved in recognition and entry into the host cell. The S protein is a type 1 membrane glycoprotein and is the main inducers in neutralising antibodies. Thus this makes it a highly important target for drug development.

The Main protease or the 3CL protease is the main target for drug development in coronaviruses including MERS and SARS. It does not have a human ortholog hence it makes it all the more important target for drug development so as not to accidentally target any human gene. The main function of the Main protease is in viral maturation by cleaving the large polyprotein 1 ab into 16 non structural proteins (nsps). The cysteine proteases work as a homodimer and mainly comprises of three domains, homodimeristation plays an important role in the catalytic activity of the protease.

The main aim of this work was to identify various hidden features of the spike protein and main protease in terms of supermotifs and centrality of amino acids by residue interaction networks. In this work we studied 208 isolates of SARS-Cov2 to identify these features in the spike protein and the main proteases.

Material and Methods

Structural retrieval

A number of recent structures are submitted in RCSB-PDB (https://www.rcsb.org/) in year 2020 of different protein domains of SARS-Cov-2(Supplementary file S1, table1). We examined all the structures submitted to the protein databank and considered protein 3D structures of given ids; 6LU7, 6LZG. All the structure were visualized in chimera tool(Pettersen, et al., 2004). The data of 219 different virus strains submitted in ncbi (https://www.ncbi.nlm.nih.gov/) were downloaded and processed for specific protein sequences

RIN and motif analysis

All the sequences were used for conserved motif finding using meme tool(Bailey, et al., 2009). The structures obtained from RCSB-PDB were used to generate the residue interaction network in cytoscape(Shannon, et al., 2003) tool (https://cytoscape.org/).

Results and discussion

Protein domain analysis

The SARS-Cov-2 is consists of 4 main structural and many non-structural proteins. These proteins play a critical role in proper functioning and replication of virus in its life-cycle and during transmission. We examined the available sequence as well as structures of the spike protein and the main protease (Supplementary File S1, Supplementary Table 1) for more specific details and hidden features. All the sequences were submitted to SMART(Letunic, et al., 2015) tool for their domain and sub domain identification. This analysis revealed the presence of multiple significant sub domains with in the main domain, contributing at sequence and structural level.

Spike protein:

The spike protein is one the major structural protein present on the surface of the SARS-Cov2. It is a type 1 membrane glycoprotein and projects on the surface of the virus like a club. It helps in attaching the virus to the host cells and is also the main inducer for neutralising the antibodies. Viruses without the spike protein loose their ability to infect the host cells. The S protein is divided into two regions, S1 and S2. S1 region consists of N-terminal NTD domain and C-terminal CTD domains (Qian, et al., 2015). The S1 domain has the receptor binding domain (RBD) that binds to

the ACE2- receptor and the S2 domain helps in the fusion of the viral and the cellular membranes, thus ensuring the entry of the virus into the cell(Ou, et al., 2020). The spike protein sequence from these 219 human strains were extracted and analysed for the identification of other conserved domains. The available data of 219 human strains from different location allowed us to examine the protein domain structure of spike-rdb region, this domain includes 4 overlapping sub-domains, which are s48_45, elicitin, WR1 and ChtBD3 (Figure. 1, Panel1).

The s48_45 domain were first observed in Plasmodium and thought to be unique to them (van Dijk, et al., 2001), however later it was observed in other aconoidasidans. They are a small family of proteins with only 14 members identified till date with each members being stage specifically expressed during the Plasmodium life cycle (Arredondo and Kappe, 2017). They are localised on the surface of the Plasmodium mostly by glycosyl phosphatydylinositol-anchoring and have been involved in the male/female gamete fusion in the midgut of the mosquito. These proteins have been targets for transmission blocking vaccine for years.

This domain shows higher number of cysteine which are positionally conserved. It forms a beta-sandwich by two sheets with parallel and antiparallel strands. These domains are possibly involved in adhesion and structurally related to ephrins, the ligand of Eph receptors and are also a key target for vaccine development against Plasmodium. It is possible in case of SARS-Cov-2 that they are using the mode of action of s48_45 for interaction/recognition with the receptor for cell invasion.

The second sub-domain elicitin, belongs to the fungal family of toxic proteins elicitins known for causing necrotic and systemic hypersensitive response in plants(Baillieul, et al., 2003), this necrosis allows immediate control of fungal invasion. They also show features of microbe-associated molecular patterns (MAMPs). Depending on their net charge these domains are divided into alpha and beta elicitins. The alpha elictins are acidic and the beta elicitins are basic and 100 times more toxic. The amino acid at position 13 is considered to cause necrosis and is speculated to be involved in ligand/receptor binding.

The third sub-domain is WR1 known as Worm-specific repeat type 1, they are cysteine rich in nature, and may have associated function with serine peptidase inhibitors. They are thought to be unique to *C. elegans*. The proteins having this domain are involved in inositol phosphate metabolism and benzoate degradation pathways. Much information is not available about them.

The fourth sub-domain was chitin-binding domain type 3 which works as carbohydrate binding module, they are involved in carbohydrate metabolic process where the function includes hydrolase

activity, hydrolysing O-glycosyl compounds and carbohydrate binding (Tomme, et al., 1988). Interestingly these sub-domains are not present in other pfam entries of spike protein domain, which indicates towards the selective presence of such sub-domain in spike domain to make recognition and binding of ACE-2 easy.

Main protease:

The main protease or Mpro is a virally encoded chymotrypsin like protease that cleaves the large polyprotein 1 ab into 16 non-structural proteins or nsps. It is one of the main target against SARS-Cov-2 in different clinical and drug designing as it does not have a human homolog. The main proteases comprises of three main domains I-III. The three domains are chymotrypsin-like 77 domain I (residues 10-99), picornavirus 3C-protease like domain II (residues 100-78 182) and helical domain III. Each domain minimally interacts with each other in homodimer formation, the majority of the interaction is between 82 residues of domain (Zhang, et al., 2020).

The domain analysis of main protease, which is a peptidase_C30 in nature (Barrett and Rawlings, 2001), revealed presence of 4 sub-domains which are LRRCT, NADH-G_4Fe-4S_3, VWC and defensin (Figure. 1, Panel1).

As it is well known that main protease is involved in viral protein processing through proteolytic cleavage of long polyprotein orf, these domains can provide very useful information about the activity of main protease. The LRRCT domain are found in a wide range of organisms from viruses to eukaryotes and provides structural framework for the formation of protein-protein interaction and are involved in a variety of biological processes, including signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, apoptosis, and the immune response (Enkhbayar, et al., 2004). This domain folds into and arc or horseshoe shape and might be involved in the structural conformation of the main protease. The NADH-G_4Fe-4S_3 domain is known for oxidoreductase activity and is found from bacterial to mammals (Lemire, 2015). This domain consists of two alpha helices separated by a loop region that brings together a [4Fe-4S] cluster through an unusual His and three Cys residue motif H-x(3)-C-x(2)-C-x(5)-C.

The third domain VWC is found in various plasma proteins and is named after the von Willebrand factor (VWF) type C repeat which is found in multidomain protein/multifunctional proteins involved in maintaining homeostasis. This domain is found in all eukaryotes and many intracellular proteins involved in transcription, DNA repair, ribosomal and membrane transport and the proteasome and functionally associated with protein binding (Hunt and Barker, 1987).

The fourth domain defensins are 2-6 kDa, cationic, microbicidal peptides active against many Gram-negative and Gram-positive bacteria, fungi, and enveloped viruses and are Cysteine-rich domains that lyse bacteria, fungi and enveloped viruses by forming multimeric membrane-spanning channels (White, et al., 1995).

The extra domains described in this section are present only in the human SARS-Cov2 and are absent in Bat Cov2 and present in pangolin with different signature sequence. This suggests that these domains that have been incorporated in the human Cov2 are important for the crossing of the zoonotic barrier and increased infectivity in the humans. In the case of spike proteins the hidden domains are important for better penetration of the virus in human cells. The Mpro is the main target for drug development it is essential to understand the protein in detail including its sub domains. In view of the known activity of the Mpro for processing the polyprotein, the sub domains are mainly involved in providing enhanced structural stability.

Figure.1

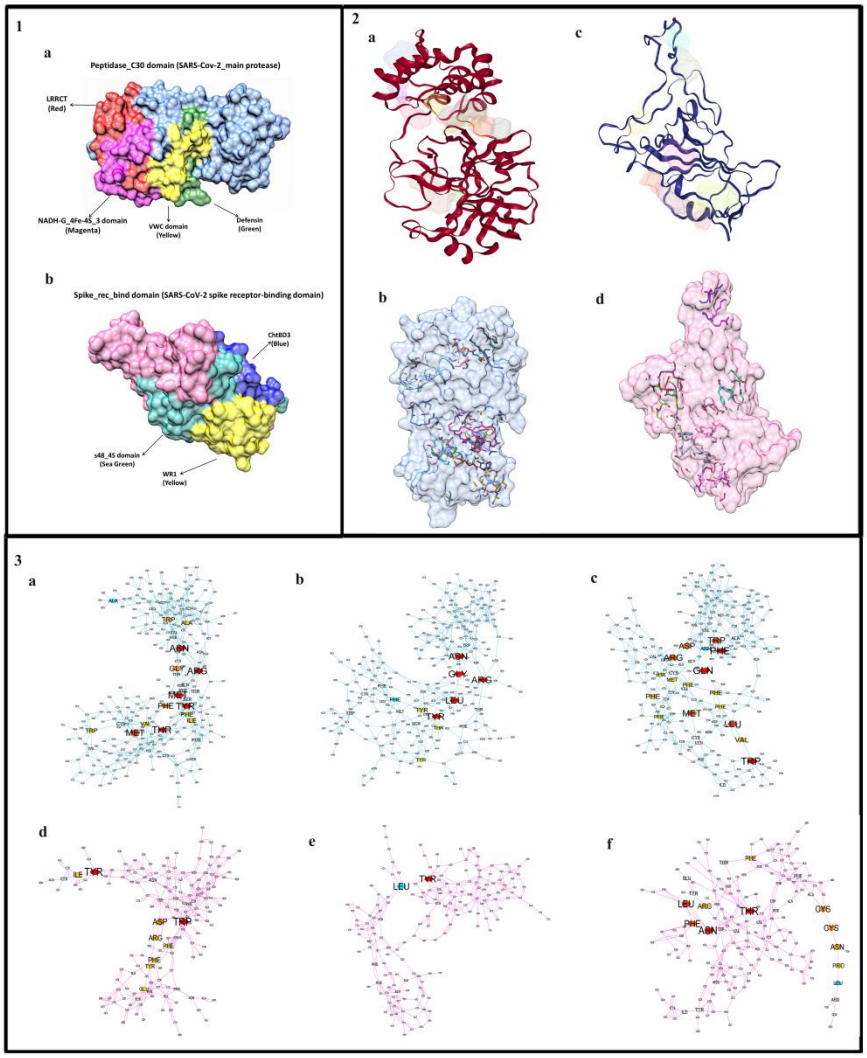

Figure 1

Panel 1: 3D structural surface representations of major subdomains present in main protease (a) and spike protein (b). **Panel 2** : main and secondary active sites in main protease (a, b) and spike protein (c, d), in mesh and surface view. **Panel 3**: pictorial representation of residue interaction networks (RIN) in 3D protein structure of main protease of human (a), bat (b) and pangolin (c) and spike protein of human (d), bat (e) and pangolin (f) based on residue centrality analysis, based on the change in average shortest path length under removal of individual nodes. The figure highlights the major residues governing the centrality in red and minor in orange and yellow, in protein structural network.

Protein active site analysis and RINs

The protein structures of the spike protein (S) and the main protease (Mpro) were analysed computationally using the tools castp (Tian, et al., 2018) and dogsitescorere (Volkamer, et al., 2012) for their active site details. The protein structures of S and Mpro were downloaded from the rscb-pdb database with the most recent structure id 6lzg (Wang, et al., 2020) and 6lu7 (Zhang, et al., 2020) respectively. Apart from this the structures were examined for orientation related details of other clefts and active site in pdbsum database (Laskowski, et al., 2018). The data in pdbsum revealed a number of secondary structural hotspots in the form of clefts and pores. The position and distribution of these structural hotspots with in major and secondary protein domain regions, clearly indicates towards their active involvement in protein function as well as in support of main functional site. It also provides a clue, that targeting the hotspots with main active site could provide and effective lead for drug development against SARS-Cov-2.

The active site and binding pocket analysis revealed that more than 100 amino acids are potentially contributing for the active site and hotspot formation in spike protein structure. The major active site is supported by Phe338, Glu340, Val341, Phe342, Asn343, Arg346, Tyr351, Arg355 and Asp364. Apart from these amino acids their a number of amino acids which are significantly supporting the structure in formation of hotspot sites (minor active sites) (Figure. 1,Panel2), (Supplementary file S1, Supplementary Table2)

The spike protein of human, bat and pangolin for their RIN (Figure. 1 Panel 3). The RIN showed that human spike needs 2 major (Trp436, Tyr489) and 4 minor (Arg355, Asp398, Asn422, Ile472) amino acids for central dependency, while spike protein of pangolin shows 4 major (Phe214, Leu217, Asn190, Thr64) amino acids and 7 minor amino acids (Arg26, Phe63, Cys196, Cys62, Asn215, Pro250, Leu253). In this comparison bat spike protein shows more resemblance with the human

spike protein where it also shows dependency on 2 major amino acids (Leu177, Tyr117), which indicate towards its conserved function (Figure. 1, panel3).

In a recent report on 3D structure of main protease protein, the binding of a substrate α-ketoamide 13a in a binding pocket located between domains I and II is shown. This pocket is contributed by a number of conserved amino acids which are His41, Met49, Cys145, Gly143, Ser144, His163, His164, Phe140, Ser1, Glu166, Pro168, and Gln189. The same protein structure was analysed computationally, for its active site details by using 2 different tools, castp and dogsitescorer (Fig. 1). The analysis showed similarity with the reported active site as well as involvement of the other important residues (Supplementary file S1,Supplementary Table3)

On the basis of above mentioned observations, the residue interaction networks (RINs)(Shcherbinin and Veselovsky, 2019) were calculated and plotted, to examine the amino acid hubs and the key amino acids which are governing the centrality of the structure. The RIN for the main protease of SARS-Cov-2 (6lu7), Bat derived main protease (4yo9) and pangolin main protease was calculated (Figure 1, Panel 3). To generate the RIN of pangolin sequences, the protein structure was computationally modelled using Phyre2 tool, these structures were further refined with molecular dynamics. The refined structures were then subjected for network generation. The RINs of main protease of human, bat and pangolin revealed interesting difference between them. The interaction network of main protease (6lu7) shows major structural dependency on 3 main amino acids (represented by red balls in figure 1) Arg131, Gln127, Phe150 and some minor amino acids(Phe150,Gly109, Trp31, Val13, Ile136, Phe185, Ala206, Trp207, Ala266) represented by orange and yellow balls in figure 1. In bat (4yo9) protein structure shows central dependency on 5 amino acids (Gly112, Tyr185, Asn206, Arg134, Leu133) with 3 minor amino acids (Thr176, Tyr164, Phe180). On the other hand the structure of pangolin derived main protease shows central dependency on 4 major amino acids (Trp3, Gln99, Phe263, and Trp179) with closely related minor amino acid (Val7, Phe131, Phe153, Phe122, Phe84, Met102, Thr107, Asn175) (Figure. 1 Panel 3). The architectural distribution of amino acids in RIN of human and pangolin main protease indicate towards the major difference in structure driven function, despite having higher sequence and structural similarity. It also suggests that main protease of human is able to maintain its structural network with lesser number of amino acids, which could be an evolutionary gain to maintain its structure stability.

Conclusion

In present study we examined the structural features of SARS-Cov-2 proteins, focusing on main protease and spike-rbd. This study revealed the presence of a number of hidden domains in spike protein and main protease of SARS-Cov2 which may be responsible for the enhanced binding and functionality of these proteins. The hotspots identified in these proteins can be helpful in drug development against SARS-Cov2. The RIN analysis will be helpful in better understanding of nature and functional mechanism of these proteins in human host, and will help in designing more targeted strategy against covid-19.

Acknowledgements

The Council of Scientific and Industrial Research, New Delhi is acknowledged for financial support to Network project (OLP-104) and senior research fellowship to YVD. University Grant Commission, New Delhi is acknowledged for senior research fellowship to PP and NB. Authors also acknowledge CSIR-4PI institute for supercomputing facility. This manuscript has been assigned the institutional id no. **CSIR-NBRI_MS/2020/05/09.**

Authors Contribution

Yogeshwar Vikram Dhar: Concept, Data curation, Formal analysis, Investigation, Methodology, Software, Writing-original draft. **Priti Prasad:** Data retrieval, Raw data preparation, Methodology. **Nikita Tiwari:** Data retrieval, Raw data preparation. **Vaishali Pankaj:** Data retrieval, Raw data preparation. **Nasreen Bano:** Data Preparation. **Sumit K Bag:** Supervision, Concept, Writing – review and editing **Mehar H Asif:** Supervision, Concept, Visualization, Writing - review and editing.

Conflict of interest

The authors declare that they have no competing interest**.**

References:

Arredondo, S.A. and Kappe, S.H.I. (2017) The s48/45 six-cysteine proteins: mediators of interaction throughout the Plasmodium life cycle, *International journal for parasitology*, **47**, 409-423.
Bailey, T.L., *et al.* (2009) MEME SUITE: tools for motif discovery and searching, *Nucleic acids research*, **37**, W202-208.
Baillieul, F., de Ruffray, P. and Kauffmann, S. (2003) Molecular cloning and biological activity of alpha-, beta-, and gamma-megaspermin, three elicitins secreted by Phytophthora megasperma H20, *Plant physiology*, **131**, 155-166.
Barrett, A.J. and Rawlings, N.D. (2001) Evolutionary lines of cysteine peptidases, *Biological chemistry*, **382**, 727-733.
Enkhbayar, P., *et al.* (2004) Structural principles of leucine-rich repeat (LRR) proteins, *Proteins*, **54**, 394-403.
Hunt, L.T. and Barker, W.C. (1987) von Willebrand factor shares a distinctive cysteine-rich domain with thrombospondin and procollagen, *Biochemical and biophysical research communications*, **144**, 876-882.
Laskowski, R.A., *et al.* (2018) PDBsum: Structural summaries of PDB entries, *Protein science : a publication of the Protein Society*, **27**, 129-134.
Lemire, B.D. (2015) Evolution of FOXRED1, an FAD-dependent oxidoreductase necessary for NADH:ubiquinone oxidoreductase (Complex I) assembly, *Biochimica et biophysica acta*, **1847**, 451-457.
Letunic, I., Doerks, T. and Bork, P. (2015) SMART: recent updates, new developments and status in 2015, *Nucleic acids research*, **43**, D257-260.
Ou, X., *et al.* (2020) Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV, *Nature communications*, **11**, 1620.
Pettersen, E.F., *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *Journal of computational chemistry*, **25**, 1605-1612.
Qian, Z., *et al.* (2015) Identification of the Receptor-Binding Domain of the Spike Glycoprotein of Human Betacoronavirus HKU1, *Journal of virology*, **89**, 8816-8827.
Shannon, P., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome research*, **13**, 2498-2504.
Shcherbinin, D. and Veselovsky, A. (2019) Analysis of Protein Structures Using Residue Interaction Networks. In Mohan, C.G. (ed), *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*. Springer International Publishing, Cham, pp. 55-69.
Tian, W., *et al.* (2018) CASTp 3.0: computed atlas of surface topography of proteins, *Nucleic acids research*, **46**, W363-W367.
Tomme, P., *et al.* (1988) Studies of the cellulolytic system of Trichoderma reesei QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis, *European journal of biochemistry*, **170**, 575-581.
van Dijk, M.R., *et al.* (2001) A central role for P48/45 in malaria parasite male gamete fertility, *Cell*, **104**, 153-164.
Volkamer, A., *et al.* (2012) DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment, *Bioinformatics (Oxford, England)*, **28**, 2074-2075.
Wang, Q., *et al.* (2020) Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2, *Cell*, **181**, 894-904.e899.
White, S.H., Wimley, W.C. and Selsted, M.E. (1995) Structure, function, and membrane integration of defensins, *Current opinion in structural biology*, **5**, 521-527.
Zhang, L., *et al.* (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors, *Science*, **368**, 409-412.