Understanding Structural Malleability of the SARS-CoV-2 Proteins and their Relation to the Comorbidities

Sagnik Sen
Jadavpur University
Ashmita Dey
Jadavpur University
Sanghamitra Bandyopadhyay
Indian Statistical Institute
Ujjwal Maulik (ujjwal.maulik@jadavpuruniversity.in)
Jadavpur University
Vladimir Uversky
University of South Florida, "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences”

Article

Keywords: SARS-CoV-2 Proteins, SARS-CoV-2, expression rates, viral protein classes

Posted Date: September 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-82352/v1

Version of Record: A version of this preprint was published at Briefings in Bioinformatics on June 18th, 2021. See the published version at https://doi.org/10.1093/bib/bbab232.

Understanding Structural Malleability of the SARS-CoV-2 Proteins and their Relation to the Comorbidities

Sagnik Sen[1,+], Ashmita Dey[1,+], Sanghamitra Bandhyopadhyay[2], Vladimir N. Uversky[3,4], and Ujjwal Maulik[1,*]

[1]Department of Computer Science and Engineering, Jadavpur University, Kolkata-32, West Bengal, India
[2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata-108, West Bengal, India
[3]Department of Molecular Medicine and Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, Florida, United States of America
[4]Laboratory of New Methods in Biology, Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", Pushchino, Moscow region, 142290 Russia 3
[*]ujjwal.maulik@jadavpuruniversity.in
[+]these authors contributed equally to this work

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a causative agent of the coronavirus disease (CoVID-19), is a part of the β-coronaviridae family. In comparison with two other members of this family of coronaviruses infecting humans (SARS-CoV and Middle East Respiratory Syndrome (MERS) CoV), SARS-CoV-2 showed the most severe effects on the entire Earth population causing world-wide CoVID-19 pandemic. SARS-CoV-2 contains five major protein classes, such as four structural proteins (Nucleocapsid (N), Membrane (M), Envelop (E), and Spike Glycoprotein (S)) and Replicase polyproteins (R), which are synthesized as two polyproteins (ORF1a and ORF1ab) that are subsequently processed into 12 nonstructural proteins by three viral proteases. All these proteins share high sequence similarity with their SARS-CoV counterparts. Due to the severity of the current situation, most of the SARS-CoV-2-related research is focused on finding therapeutic solutions and the analysis of comorbidities during infection. However, studies on the peculiarities of the amino acid sequences of viral protein classes and their structure space analysis throughout the evolutionary time-frame are limited. At the same time, due to their structural malleability, viral proteins can be directly or indirectly associated with the dysfunctionality of the host cell proteins, which may lead to comorbidities during the infection and at the post infection stage. To fill these gaps, we conducted the evolutionary sequence-structure analysis of the viral protein classes to evaluate the rate of their evolutionary malleability. We also looked at the intrinsic disorder propensities of these viral proteins and confirmed that although they typically do not have long intrinsically disordered regions (IDRs), all of them have at least some levels of intrinsic disorder. Furthermore, short IDRs found in viral proteins are extremely effective and prioritize the proteins for host cell interactions, which may lead to host cell dysfunction. Next, the associations of viral proteins with the host cell proteins were studied, and a list of diseases which are associated with such host cell proteins was developed. Other than the usual set of diseases, we have identified some maladies, which may happen after the recovery from the infections. Comparison of the expression rates of the host cell proteins during the diseases suggested the existence of two distinct classes. First class includes proteins, which are directly associated with certain sets of diseases, where they have shared similar activities. Second class is related to the cytokine storm-mediated pro-inflammation (already known for its role in acute respiratory distress syndrome, ARDS), and neuroinflammation may trigger some of the neurological malignancies and neurodegenerative and neuropsychiatric diseases. Finally, since the transmembrane serine protease 2 (TMPRSS2), which is one of the leading proteins associated with the viral uptake, is an androgen-mediated protein, our study suggested that males and postmenopausal females can be more susceptible to the SARS-CoV-2 infection.

Introduction

From coronaviridae family, three types of β-coronaviruses, such as Middle-East respiratory syndrome coronavirus (MERS-CoV), severe acute respiratory syndrome coronavirus (SARS-CoV), and SARS-CoV-2, are able to infect humans causing fatal pneumonia. Chronologically, SARS-CoV and MERS-CoV were the first such hCoVs, causing death of 774 (in the year of 2002) and 885 (in the year of 2012), respectively. SARS-CoV-2 causing current CoVID-19 pandemic has surpassed all the previous records of the hCoV infection with more than 25.0 million infection and more than 850,000 death cases in countries

and territories around the world and 2 international conveyances. Amino acid sequences of the SARS-CoV-2 proteins are sharing high similarity with SARS-CoV proteins (89% on average) (see Figure 1). SARS-CoV-2 proteome consists of four unique classes of proteins that provide the structural basis of the virus and protect its positive single-stranded RNA genome (+ssRNA). On the other hand, replicase polyproteins, which are synthesized as two polyproteins (ORF1a and ORF1ab) that are subsequently processed into 12 nonstructural proteins by three viral proteases ORF1ab polyproteins, play important roles in regulation of the +ssRNA replication within the host cells. The entry in the host cell is started via interaction of the Spike glycoprotein (S protein), from the outermost membrane, with the Angiotensin Converting Enzyme 2 (ACE2) receptors on the surface of some host cells. After that, the replicases are responsible for the interfering with various host cell mechanisms. Also, the Membrane (M) and Nucleocapsid (N) proteins have their own set of distinct functions, which involves interactions with lots of available host cell proteins. As per previously published studies, some of these viral proteins have long intrinsically disordered regions (IDRs). For example, in[1], high level of disorder in the nucleocapsid from SARS-CoV was shown. Similarly, the viral M proteins are known for the presence of functionally important IDRs. In a recent review, the predisposition of viral envelop proteins for intrinsic disorder was systemized[2]. Almost each protein from the SARS-CoV-2 has previously been shown to contain functional IDRs[3]. In a recent study, N, M, and E proteins for SARS-CoV-2 have been studied for their structural disorder[4]. Therefore, the presence of structural flexibility may have an impact on interactions of viral proteins with host cell partners. This can be one of the reasons behind the vulnerability of the infected patients with comorbidities.

It has been observed that comorbidities play a crucial role in the rate of fatality during COVID-19. Although SARS-CoV-2 seems to be characterized by multiorgan tropism[5], the reasons behind this phenomenon are poorly understood. Apart from the injury risks of some vital organ, such as lung injury[6], kidney failure[7], liver damage[8], and heart diseases[9,10], the association of the SARS-CoV-2 proteins with the host cell proteins can be observed in diverse conditions. Along with that, the cytokine storm is an experimentally observed phenomenon during SARS-CoV-2 infection[11]. Following this path, the connection between viral pathogenesis and immunopathological elements can also be observed, where some of the vital pathways, e.g., NFkB signaling pathways are also found to be differentially regulated[12]. The cross talk between the host cell proteins and viral proteins might have promote certain condition. Actually, multi-valency of the protein can increase the possibility of multi interaction partners. In a recent study, high binding potential of the IDPs has been established[13]. This explains how the structural malleability and the multi-valency of the viral protein may have prioritized the interaction with host cell proteins and regulates the host cell pathways.

Almost each of the SARS-CoV-2 proteins has been intensively studied for its druggability. In Wu et al.[14], Spike Protein and Replicases, e.g., RNA-dependent RNA polymerase (RdRp), papain like protease (PLpro), etc. were analyzed as potential drug targets for SARS-CoV-2. Similarly, N protein is expected to be a prime candidate for its antigenicity. However, very few of such drug discovery studies have mentioned the structural malleability and intrinsic disorder potential of these proteins. In Xue et al.[15], the trait of structural disorder for the viral proteins was described. Similarly, the predisposition for structural disorder and potential consequences of the presence of IDRs were pointed out for M and N proteins in light of the discussion of the rigidity of the outer shell of SARS-CoV-2[16]. It is clear that studies of each category of the viral proteins may provide insights into their structural instability. This will help in better understanding of the functional roles of their flexibility for interaction with the host cell partners, e.g., ACE2 as a receptor of Spike S1 glycoproteins. In this article, we have aimed to

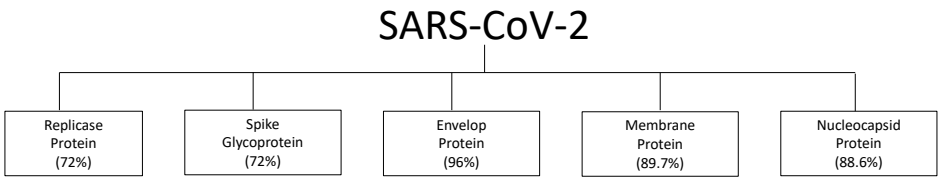

Figure 1. Sequence similarity of the selected proteins of SARS-CoV-2 and SARS-CoV

study structural flexibility of the SARS-CoV-2 proteins and comorbidities due to coronavirus infection. As per previous studies, many host systems may become dysfunctional as a result of viral infection, and such vulnerability of the natural host cell systems can be determined by structural malleability of the viral proteins. Therefore, we have started with the analysis of the evolutionary sequence-structure space[17–19] of major SARS-CoV-2 proteins. This study helps unveiling the mutational sequence landscape of each class of viral proteins based on the residue occupancy[20]. Subsequently, evolutionary coupling propensity[21] has been calculated, and the corresponding sequence space networks[22] have been designed depending on each protein family information. In the structure space, the individual structural root mean square-fluctuations have been studied. After that, the structure network was created for all the viral proteins. These information can provide important insights on structural flexibility of viral proteins. The viral proteins have defined activities during the cell entry as well as within the host cell. As per their host

cell targets (both interaction partners and pathways), each of the viral protein participants are partially associated with diverse set of diseases. We have grouped the diseases within networks, considering the protein-protein interaction (PPI) networks of the host cell participants. The resultant outcomes have emphasized the link between the flexibility of the viral proteins and SARS-CoV-2 associated comorbidities. Hence, the patients with certain sets of pathological conditions are expected to be more vulnerable than patients without comorbidities. Also the risk of diverse diseases which are not pre-determined, can be increased due to unusual dysfunction within the host cells.

Results

The aim of the study was to reveal the structural flexibility of the viral proteins associated with SARS-CoV-2 infection. Furthermore, the diseases accompanying COVID-19 were also studied in details. It is evident from the reported results that high flexibility of the selected proteins is responsible for the vulnerability of population to this infection. The study is performed in two different stages. In the first stage, the proteins are analyzed based on their sequence and structure space, whereas in the second stage, the disease associations are established.

A hydropathy plots generated for each protein sequence belonging to 17 protein families are shown in Figure 3. Such hydropathy plots allow for the visualization of hydrophobicity over the length of a peptide sequence. A sliding "window" determines the summed hydropathy at each point in the sequence (Y coordinate). These sums are then plotted against their respective positions (X coordinate). Such plots are useful in determining the hydrophobic patterns of globular proteins, as well as determining membrane spanning regions of membrane the proteins. subsectionSequence Space Analysis From Pfam, we

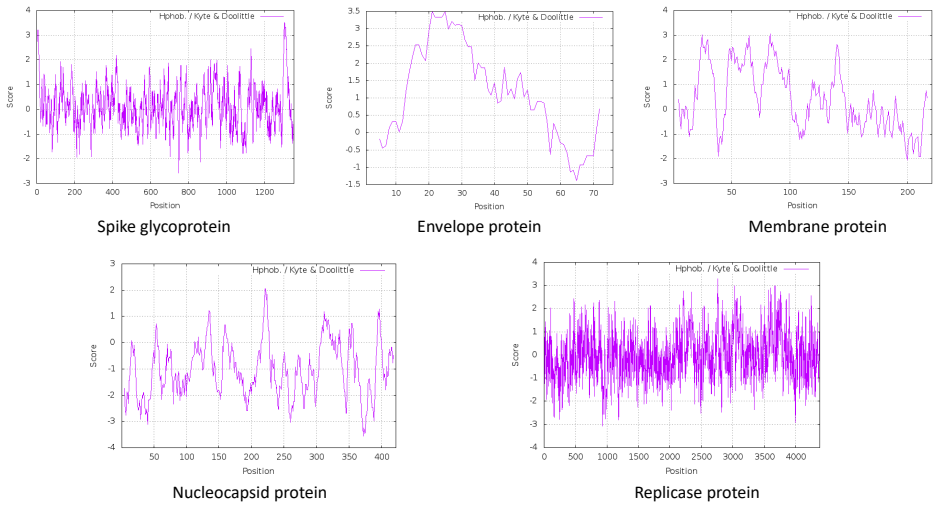

Figure 2. The two-dimensional hydropathy plot produced for the selected proteins based on the Kyte and Doolittle scale of amino acid hydropathy.

have selected 17 protein families consisting at least one viral protein. Among them, few proteins, at least one from each family are reported in the main paper and rest of the proteins are given in supplementary Figure S1. During the experiment, we have faced one such consequences, where a protein can be a member of more than one protein family (homologically). In such cases, we have considered the most appropriate one, e.g., PF01600 is more appropriate family for Spike glycoproteins. Similarly, Nsp1 is also associated with two families based on the sequence similarity at nucleotide-binding domain and polypeptide conservation. After the selection, SE scorings have been provided for each family. In this case, column-wise sequential occupancy has been reflected in the SE scoring. As aforementioned, we have started with individual protein families. Now, the sequential columns of each family are considered as evolutionary stable at residual position if the score is zero. Higher scores represent more sequential randomness. Therefore, the protein specific SE plots in Figure 4 represent the degree of their evolutionary randomness. Simultaneously, the coupled pairs are studied based on distinct scoring i.e., MI and DI. The co-occurrences of the residues have been shown in the feature map with MI scores. The graphs of each protein show some small white dots which represent the conserved co-varying patches. These also helps portraying the coupling strength between the residual position or the co-varying amino acids. In the case of DI, we have designed weighted networks from the DI scores and corresponding coupled pairs. Hence, the network can represent the evolutionary conserved inner allosteric. The communities from the network are based on the eigenvector centrality. Each community has at least one node with higher eigenvector centrality scores. The

connected nodes are involved and responsible for the corresponding scores. Therefore, eigenvector-based communities can represent the internal modifications due to the mutational changes at any residual nodes from the network.

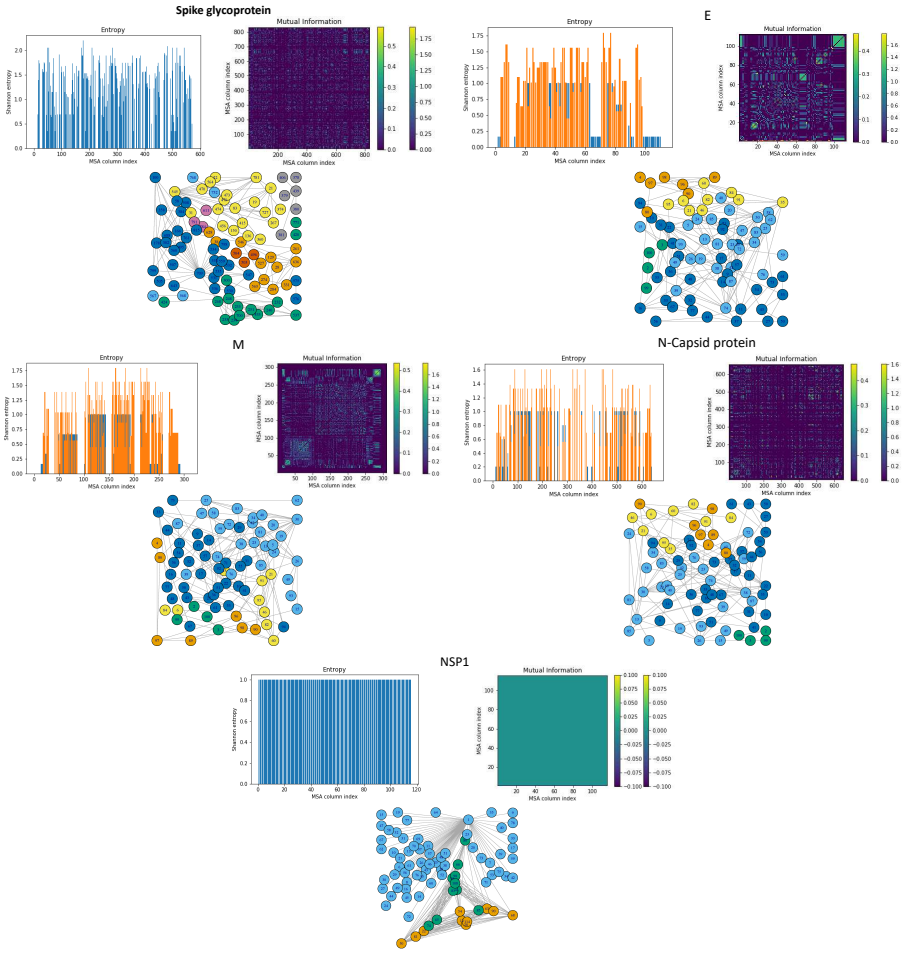

Figure 3. The sequence space analysis is performed for the selected protein families along with the Shannon Entropy calculation, coupling analysis, and community detection techniques

Structure Space Analysis

In Figure 5, the square-fluctuations of the protein sequences is analyzed by utilizing eigenvector centrality. These fluctuations depict the mutational link between sequence and structure of a particular protein. The peak of fluctuating regions is mapped with the conserved region of the protein by using the sequence space knowledge. This decipher how the conserved regions are associated with the mutational changes, which may lead to the change in structure-function association. Furthermore, depending on the DI scores, the residues in the same cluster represent the same rate of evolutionary change. In parallel, to decipher the residue-wise organization and dependencies at different secondary structure elements, structure network analysis was employed. The potential 3D structures of these proteins are modelled using I-TASSER[23,24]. The eigenvector centrality is calculated to unveil the influence of a particular node on the internal dynamics of different protein structures. The structure network analysis is performed to map the sequence space changes in structure space shown in Figure 5.

Diseases Related to the SARS-CoV-2 Infection

Current literature contains a wealth of information supporting the notion that the patients suffering from cardiovascular diseases, gastrointestinal disorder, diabetics and cancer show higher vulnerability towards the SARS-CoV-2 infection. Patients in critical condition are more prone to this disease due to the multiple organ failure. Unfortunately, no particular therapeutic means or treatment are found that can cure the affected patients. Furthermore, diverse pathways and biological process are affected due to the infection, which leads to comorbidity. In order to understand the effect of this infection, in the second stage of this study, the association of diseases with CoVID-19 was considered. Firstly, the relations between the SARS-CoV-2 proteins and

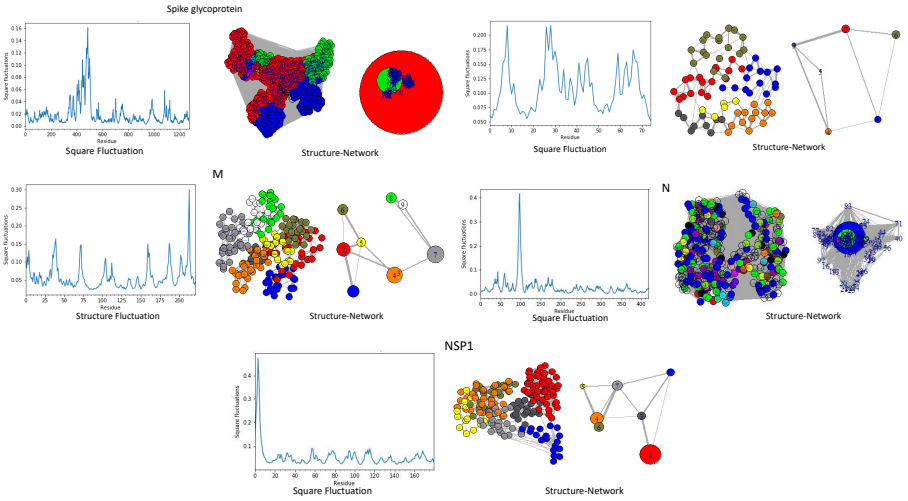

Figure 4. In the structure space analysis of the selected protein families the structure fluctuations and structure network are performed.

the host proteins were established. From that relationships we have identified 14 human proteins showing crucial impact on specific disease. Using STRING database, 14 individual PPI networks were constructed. During this network construction, the neighboring members of the PPI network are selected based on the sharing same protein homology, experimentally curated and same co-expression. For each selected protein and their neighbors, all possible diseases are curated from the DisGeNET[25]. The list of diseases was compared with the shortlisted disease list mentioned in the Method section.
In Figure 6, ten diseases are reported, and they are selected based on the number of proteins having an impact on it. During SARS-CoV-2 infection, multiple body organ are affected, such as lungs, kidney, liver, cardiovascular system, etc. To understand the involvement of the proteins into the distortion of the organ behavior we constructed a Venn diagram of the diseases shared by the proteins. Interestingly, this analysis revealed that diverse diseases have resulted from the misbehavior these proteins. Among multiple diseases, ten diseases are reported in a bar plot along with the number of associated proteins in this paper. Furthermore, the dependency score of the diseases with their respective protein are considered, and a corresponding line graph is constructed (see Figure 7). Each line of the graph represents a particular disease marked in the figure. The reported diseases include those characterized by a high risk of organ failure and others showing an impact on COVID-19.

Those common diseases are selected and a network is built between the proteins and their resulting diseases. As particular organs are highly affected due to this disease, we have categorized the diseases according to specific organ and represent them with diverse colors. In Figure 8, five proteins (PHB, FURIN, TMPRSS2, ACE2, SMAD3) are shown. The proteins are selected by analyzing the Venn-diagram based on the maximum common diseases among 14 proteins. Rest of the protein and their associated disease networks are reported in supplementary Figure S2. In the network, oval- and square-shaped nodes represent proteins and diseases, respectively. On the other hand light pink, deep blue, light green, deep green, violet, grey, c-green, orange, and light blue colors of nodes represent different diseases related to COVID-19, such as cardiovascular disease, mental health, immune system diseases, respiratory disease, cancer, diabetics, kidney disease, gastrointestinal disease and others, respectively.

Discussion

Evolutionary Sequence-Structure Space Study of Proteins E, M, and N

Many membrane proteins are known to have biologically important IDRs. Large conformational flexibility of those regions might be related to multifunctionality of these proteins[26]. Among the SARS-CoV-2 proteins, the M and E proteins are quintessential membrane proteins. Figures 9A and 9B represent the intrinsic disorder profiles generated for these proteins using several commonly used per-residue disorder predictors and show that although both M and E proteins are mostly ordered, they are expected to have disordered N- and C-terminal regions. This is also in line with the results of the hydropathy analysis of these protein (see Figure 3), which shows that these proteins are enriched in hydrophobic residues (i.e., residues with positive hydropathy values in Kyle-Doolittle scale). This enrichment in hydrophobic residues is a typical hallmark of transmembrane proteins. On the other hand, N- and C-tails of both M and E proteins are enriched in hydrophilic, polar amino acid. In this analysis, outputs are per residue predictions of the intrinsic disorder propensity in the [0, 1] range. These outputs are then compared to a threshold (we used the default threshold 0.5), and residues with a prediction value greater than the threshold

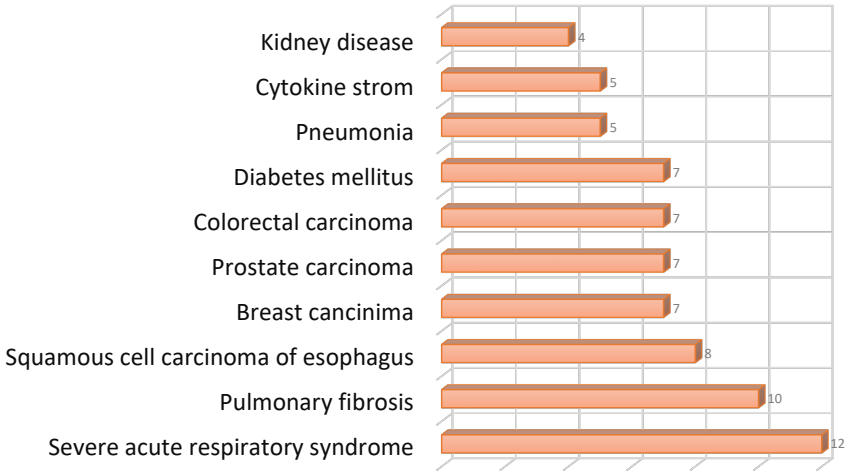

Figure 5. To understand the involvement of the proteins with the distortion of the organ behavior we performed a Venn diagram of the disease shared by the proteins.

were predicted to be intrinsically disordered. The local increase in the intrinsic disorder tendency for an ordered protein with overall low disorder propensity scores is expected to correlate with the increased mobility of the studied region[27–32]. In SARS-CoV, the functions of the M and E proteins are diverse and enigmatic. The higher sequential similarity enhances the possibility of physiological similarity for these proteins in SARS-CoV-2. In case of M protein, the assemblies with three distinct classes provides a specific set of functions. Specifically, E and M are involved in intercellular trafficking in Endoplasmic Reticulum (ER), Golgi Apparatus, and ER-Golgi intermediate compartment (ERGIC) after host cell intrusion. Mechanistic advantages have been attributed to structural flexibility of some of the parts of these proteins, which may provide the higher adaptation ability at diverse condition (at least more than the soluble proteins)[39]. In Figure 4, the residual occupancies have been highlighted for both the proteins. The conservation within the N- and C-terminal regions is quite strong. However, intermediate part is extremely susceptible to mutations in the E protein, whereas for M protein, intermediate part is slightly conserved. On other hand, the coupling analyses have shown a different aspect based on evolutionary co-variation or co-occurrence. In MI, the distributions of the coupled pairs are analyzed within the sequence space of E and M proteins. However, DI shows slightly different aspect. Firstly, DI considers only the coupling propensities (or co-occurrence trait) between two residues in a disentangled manner. The coupling score-based weighted network has a minimal number of residues. Subsequently, application of the eigenvector-based compartment distribution clarifies the evolutionary cross-talk between residues, which are extremely conserved within few color modules. In E proteins, most of the residues are distributed within two color modules; i.e., cyan and blue, where the residual range is within 100. However, the residual occupancies are evolutionary random at those sequential positions. As mentioned earlier, DI weighted network is quite different for M proteins. Interestingly, M protein network has distinct modules, which are highly aligned with SE scores. More elaborately, highly random sequential columns belong to the same colored module. The longest conserved region as per SE scores has kept within the blue colored module. It emphasizes an important point that the rate of flexibility may enhance the rate of adaptability. It is quite evident, as E and M proteins are basically sub-classes of membrane-bound proteins. The amino acid sequences of the M proteins are extremely variable. In homologically similar viruses, the disorder rate of M proteins is within 4%-14% at least. However, not much of information about E proteins is known. In the structure space study, the square-fluctuations of E and M proteins are almost similar. However, the residues at structure network of E and M proteins are distributed in such a way that even a small percentage of disorder can affect the whole structure. In the case of M proteins, the C-terminal residues are conserved with transmembrane residues with two structure network modules. In the case of E proteins, the two bigger modules include maximum number of residues, which covers almost the whole structure (all shown in Figure 5). Therefore, the sequence vulnerability can provide the structural adaptability at diverse condition.

Similarly, N proteins, known for their RNA binding property, are also being studied in the similar manner. In[40], it has been shown that the disorder at binding zone for nucleotide binding protein is quite obvious. In fact, a systematic computational analysis of the nucleiomes (∼548 000 nucleic acid binding proteins) of 1121 species from Archaea, Bacteria and Eukaryota revealed the prevalence of intrinsic disorder in these proteins[41]. Based on the analysis of 5658 dissimilar (below 50% sequence similarity) proteins with known 3D-structures that bind to proteins, DNA or RNAs it was also pointed out that disorder is crucial for the formation of many protein–protein, protein–DNA and protein–RNA complexes, where IDRs undergo the

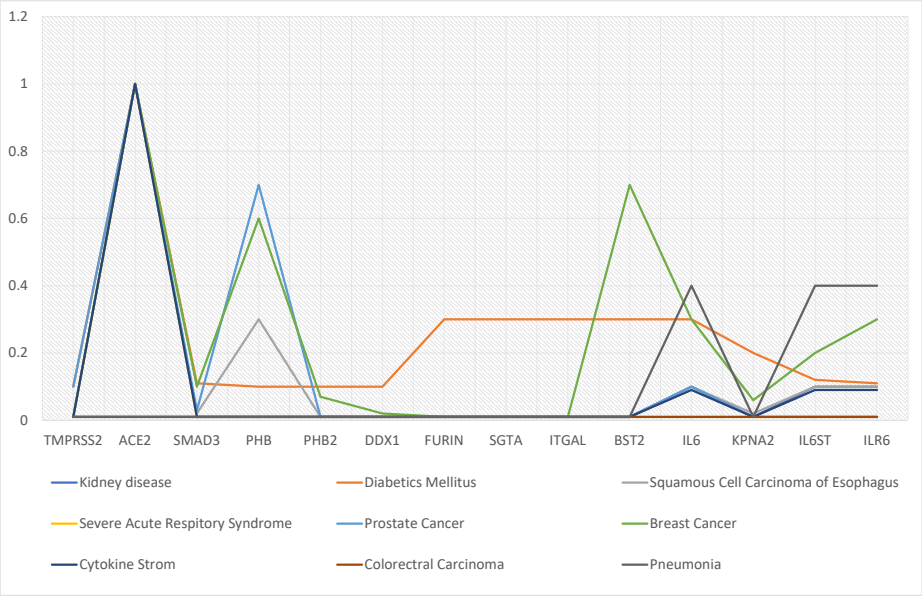

Figure 6. The dependency score of the diseases with their respective protein are considered and a line graph is constructed. Each line of the graph represents a particular disease marked in the figure.

binding-induced disorder-to-order transitions[42]. Also, the significant rate of disorder has been predicted for N proteins in[43]. Interestingly, N- and C-terminal regions of N proteins are IDRs (see Figure 9C). However, we are expecting that IDRs at N proteins might have strong impact on the receptor binding domain (RBD). Initially, the N proteins bind with viral RNA genome. It is also involved in the regulation of the replication cycle and is involved in the host cellular response to viral infection, which finally leads to the formation of virus like particles (VLPs). Usually, DNA and RNA binding proteins are known for their lower hydrophobicity. In Figure 3, the hydropathy plot provides a clear support to this statement. In Figure 4, the residual occupancy of N proteins shows a random distribution of the SE score throughout the sequence space, where the positions of some of the resides are locally conserved. Interestingly, coupling propensities of the N proteins are almost similar to those of the M proteins. In MI map, MI patches are less distributed than in E and M Proteins. However, the communities of weighted network based on DI scores is equally distributed within the residues. Interestingly, some of the residues from each module are from the conserved portion of the SE score distributions. Therefore, the sequence space is extremely vulnerable and random. In[44], the structural disordered rate of the homological similar viruses for N proteins is around 48%-57%. Therefore, IDRs at the N- and C-terminal regions of the N protein can affect the RBD and nucleotide binding domains. These also reflect in structure network shown in Figure 5. In the structure network, almost all the residues are conserved within blue modules. The previously analyzed information in terms of sequence space and structure space clarifies the trait of unfolding scenario at monomeric stage.

Evolutionary Sequence-Structure Space Study of the Spike S Glycoprotein

For S proteins, not much information has been provided by the previous researchers. As per some of the information from previous researches, the spike proteins are expected to have a lesser disorder. However, structurally, spike glycoproteins are full of α-helices. Therefore, it raises the questions on the possibility of the high prevalence of structural disorder in this protein. In agreement with this notion, Figure 9D shows that S protein is predicted to be mostly ordered, but still contains several short IDRs. In Figure 3, the distribution of the hydrophobic index is shown. The lower levels of hydrophobicity can enhance the chances of local unfolding of a protein at changes in the environmental condition. Figure 4 shows that the residual occupancy is extremely random throughout the entire space, with conserved regions being extremely small. However, N- and C-terminal regions are sequentially conserved. In MI map, distribution of the coupling patches is low mainly at the tail of the sequential distribution. Interestingly, the DI-based residues and cross-talking through the communities consist mostly the tail end residues. Figure 5 shows that the square-fluctuation is comparatively low. It may possible be due to the length of the sequence, because we have considered the normalized square-fluctuations here. However, the community distribution in the structure network clarifies its inability to achieve the stability stage at monomeric phase. Almost 620 (which is more than 50% of the sequence length) residues including the RBD are grouped within the red module. Therefore, the spike glycoprotein can also be an IDP with greater flexibility.

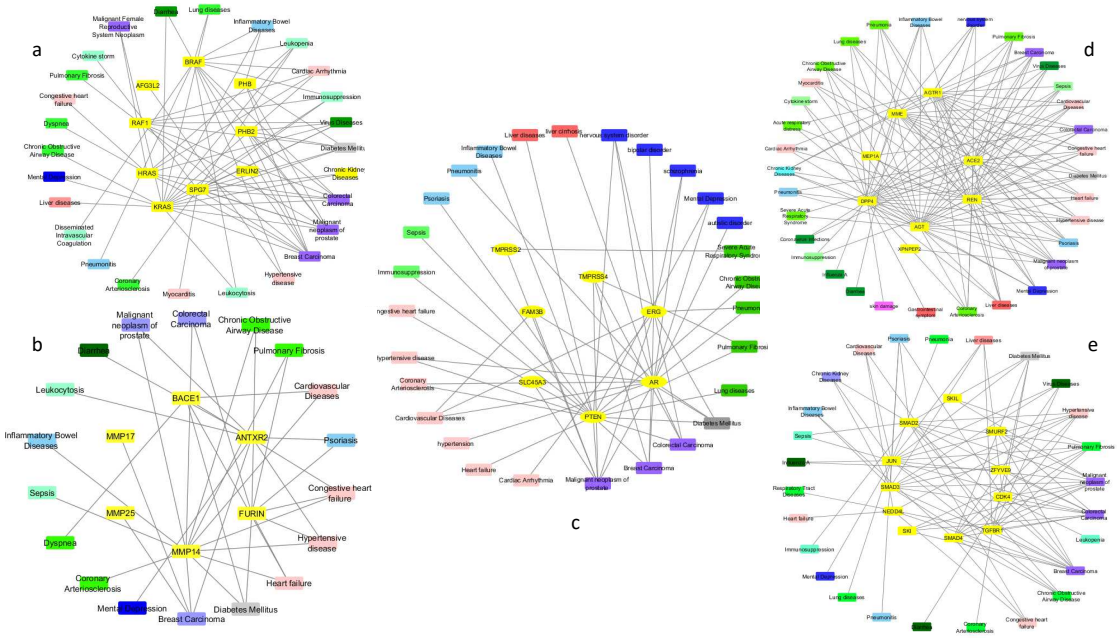

Figure 7. Five proteins a. PHB , b. FURIN , c. TMPRSS2, d. ACE2, e. SMAD3 are shown based on a common disease among the 14 proteins.

Evolutionary Sequence-Structure Space of the Replicase Polyprotein ORF1ab

Replicase polyproteins include a set of proteins, which are firmly associated with replication of the viral RNA. Individually, they are performing distinct functions. After viral uptake, replication of nucleotide is the vital role, where a large number of the host cell proteins, such as Prohibitin and Prohibitin 2, are engaged. During this process, the vital cellular organelles, such as mitochondria and ribosomes are participating through the viral interactor counterparts. Basically, the process is initiated by the host cell mRNA degradation. The complex of Nsp1 and 40S ribosome creates endonucleolytic cleavage near the 5' UTR of the host cell mRNAs, which assists the mRNA degradation process. However, leader sequence at 5' UTR of the viral mRNA protects it from the action of the Nsp1-40S complex[45]. Likewise, Nsp2 is interacting with Prohibitin and Prohibitin 2 to initiate the cell survival signals[46]. Each viral protein derived from the polyproteins is performing a distinct set of functions, and such multifunctionality can be attributed to structural flexibility of these proteins. Figures 3, 4 and 5 represents the results of the detailed sequential and structural analysis of this protein. We have only shown the results for Nsp1. As per hydropathy index, the distribution of the hydrophobic and hydrophilic residues within this is rather random. Therefore, presence of a continuous disordered regions cannot be observed throughout the sequences. Figures 9E and 9F are in line with the results of the previously published studies, where the sequences of proteins derived from the polyproteins ORF1a and ORF1ab were shown to have large sets of structurally ordered patches. However, all these proteins are expected to have some small sequence patches of disorder[3]. Interestingly, IDRs are effective enough to participate in the functional and dysfunctional protein interactions. Therefore, we have aimed to observe the fluctuating hubs of the replicase polyproteins. As discussed earlier, the hydropathy distribution is more variable. Similarly, the SE score distribution over the MSAs shows high entropy through the evolutionary sequence space. It shows that the position of each residue contains higher rate of substitution frequencies. Interestingly, sequential co-variation study has shown that the long sequential dependencies are conserved within the one color module; i.e., the cyan module. However, square-fluctuation shows a higher rate at the N-terminal end of the structure, whereas the structure network shows localized structural communities. From these results, the dependency of each residue position is clear, which also shows that disorder at any residual point can have strong impact in rest of the structure. Interestingly, other proteins (given in Supplementary Figure S3) show rather similar trait.

Host Proteins and Corresponding Viral Protein

At least 15 host cell proteins were shown to be activated during the virus intrusion. All of these are directly or indirectly affected by viral proteins and are involved in direct or indirect interactions with them. ACE2 and TMPRSS2[47,48] are two well-known host proteins, which govern the initial virus entry through spike glycoproteins. In this case, we have considered host cell proteins, which that are interacting with or affected by SARS-CoV-2 proteins, and information about which is available in UniProt. In previous focused studies, these selected human proteins were categorized as per their involvement in viral infection.

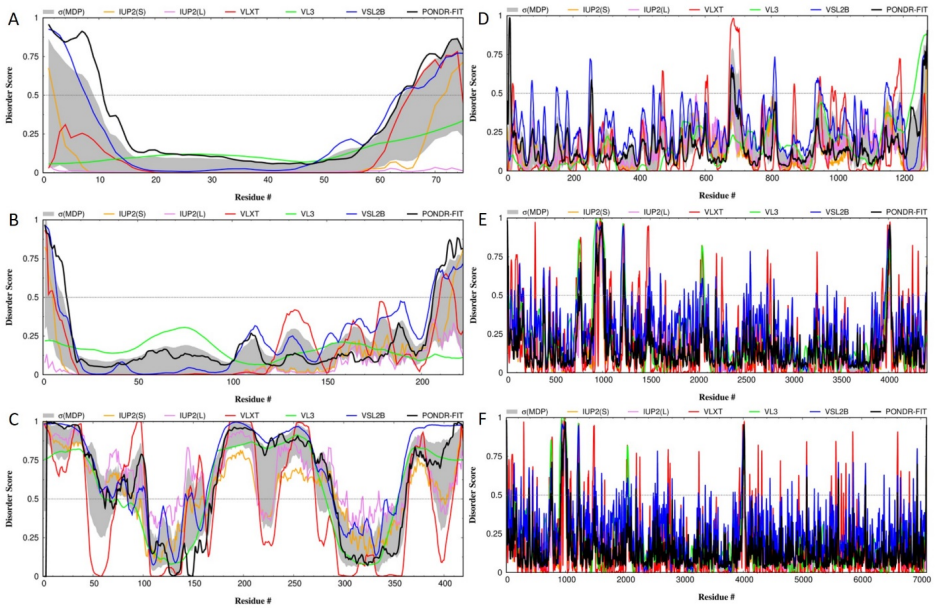

Figure 8. Evaluation of intrinsic disorder predisposition of major SARS-CoV proteins: envelope (A), membrane (B), nucleocapsid (C), spike (D), ORF1a (E) and ORF1ab (F). These profiles were generated using DiSpi web crawler that aggregate the results from a number of well-known disorder predictors: PONDR® VLXT[33], PONDR® VSL2[34], PONDR® VL3[35], IUPred short and IUPred long[36,37], and PONDR® FIT[38]

After viral uptake into the host cell, the viral proteins firstly attenuate or stop the regular protein functions. Subsequently, the cell cycle oscillation is used to grow and created multiple copies. Interestingly, not all human proteins have performed to enhance the efficiency of infection and increase the rate of virus production in the host cells. Some of these proteins attenuate the proteins, which are involved in the pathogenic progression. RNA helicase-DDX1 complex, clearly observed in SARS-CoV, is one such examples, which is related to the defense responses against the virus infection[49]. Basically, a large number of host cell proteins are involved in the viral genome replication through subparts of the replicase polyproteins. ORF1ab replicase polyprotein is a long polypeptide (7,096 residues; UniProt ID: P0DTD1) that includes multiple proteins with various activities. Specific enzymatic cleavages of this polyprotein in vivo by its own proteases, 3CLPRO and PLPRO proteinases, which are autocatalytically processed, yield mature proteins. From the N- to C-terminus of the ORF1ab polyprotein, these functional proteins are Host translation inhibitor Nsp1 (residues 1 – 180), Non-structural protein 2 (Nsp2, residues 181 – 818), Nsp3 (residues 819 – 2763), Nsp4 (residues 2764 – 3263), Nsp5 or 3C-like proteinase (3CLPRO, residues 3264 – 3569), Nsp6 (residues 3570 – 3859), Nsp7 (residues 3860 – 3942), Nsp8 (residues 3943 – 4140), Nsp9 (residues 4141 – 4253), Nsp10 (residues 4254 – 4392), RNA-directed RNA polymerase (residues 4393 – 5324), Helicase (residues 5325 – 5925), Proofreading exoribonuclease (residues 5926 – 6452), Uridylate-specific endoribonuclease (residues 6453 – 6798), and 2'-O-methyltransferase (residues 6799 – 7096). Replicase polyprotein ORF1a is a 4,405-residue-long polypeptide (UniProt ID: P0DTC1) that is produced by the ribosomal frameshifting and includes the same set of proteins found within the first half of ORF1ab (residues 1-4392); i.e., Nsp1-4, 3CLPRO, and Nsp6-10, but instead of a set of replication-related enzymes found in the C-terminal part of ORF1ab, ORF1a contains just one protein Nsp10 (residues 4393 – 4405). Starting with Nsp2[50], each of the protein excised from the viral polyproteins is performing its specific functions and is playing its specific roles in assisting the viral genome replication, modulating the cell growth, and/or controlling the pathogenic intrusion into the nucleus. As it was already pointed out, Nsp1 forms a complex with 40S ribosome, and this complex initiates the process of viral replication by inhibiting the host cell mRNAs. This complex binds at the 5' UTR of the host mRNAs and promotes their degradation. However, the viral mRNAs are not degraded, due to the presence of a specific 5'-leader sequences. Subsequently, Nsp2 interacts with host PHB and PHB2 proteins and sends the survival signal, which helps keeping the normal cell machinery and metabolism active. Likewise, each other Nsps has specific set of roles associated with interaction with the particular host cell proteins. Interesting example is given by FURIN, which shows almost similar type of proteolytic behavior as TMPRSS2[51]. In case of TMPRSS2, the viral entry mechanism has been facilitated in two ways, e.g., proteolytic cleavage of the host receptor ACE2 and proteolytic cleavage at the s1/s2 domain boundary of spike glycoprotein[47]. However, furin utilizes a proteolytic cleavage site at the spike glycoproteins[52]. SARS-CoV papain-like protease (PLpro) induces Egr-1 dependent up-regulation of the transforming growth

factor-β1 (TGF-β1) through the ROS/p38/MAPK/STAT3 pathway, thereby up-regulating the pro-fibrotic responses in vitro and in vivo[53]. SARS-CoV PLpro also caused the change in the ubiquitination profile of Rho GTPase family proteins and plays a role in the TGF-β1-dependent expression of Type I collagen via activating STAT6 pathway[54]. Also, SARS-CoV nucleocapsid (N) protein potentiates TGF-β1-induced expression of plasminogen activator inhibitor-1 and attenuates the Smad3/Smad4-mediated apoptosis of human peripheral lung epithelial cells[55]. Many other host cell proteins either work to assist the viral uptake or are affected due to viral uptake and corresponding pathogenesis. Therefore, viral proteins are capable of modulation of numerous host cell proteins. Therefore, it is possible that such virus-induced distortion in the normal proteostasis within the host cells can be associated with certain comorbidities found in recovered patients at post-COVID scenario. The list of host cell proteins associated with specific diseases has been provided.

Comorbidities and SARS-CoV-2

In continuation from the previous part, we have analyzed each of the host cell proteins for their association with diseases. In some of the previous studies, researchers have created the distinct classification in terms of associated diseases, where they have mostly studied the diseases during the infection and associated comorbidities. Apart from regular set of comorbidities (e.g., ARDS, Diabetes, chronic liver or kidney diseases etc.), several more diseases, such as malignancies and neurological disorders have been observed in infected patients. In[56], a group of patients has been observed of raising the risk schizophrenia. Another paper showed the set of infected male patients, having prostate carcinoma. We are expecting that the developing pathogenic conditions might have a strong connection with dysfunctional host cell proteins. The disordered nature of viral proteins makes them extremely sensitive to subtle changes in the environmental conditions, but also prioritizes them as promiscuous interaction partners. Therefore, the different class of viral proteins may modulate the corresponding host cell proteins (some of the examples are given in the previous section). In this regard, we have searched all the diseases associated with dysfunctional proteins. Among such associated diseases (or comorbidities) are the possibilities for co-infection, e.g., with Influenza A virus. However, the main effect of SARS-CoV-2 pathogenesis is related to the 'cytokine storm' theory. In[57], the importance of IL6 has clearly been described. Following the 'cytokine storm' theory[3], two possibilities has been observed. First, acute inflammatory activities are attributed to the increased levels of many the cytokines and chemokines, and specially IL6, with the level of IL6 being used as a predictor of the infection severity. Second, this 'cytokine storm' in Central Nervous Systems (CNS) can trigger neuroinflammation. Apart from acute lung or cardiovascular diseases, two distinct large classes of diseases can be initiated through 'cytokine storm', especially due to the increased IL6 levels. Associated diseases are more elaborately described below.

Malignancies and SARS-CoV-2 Infection

Along with other common diseases, such as hypertension, diabetes type I and type II, malignancies can be among of the vital comorbidities of COVID-19. It has been shown that malignant patients are more vulnerable to infection than the non-malignant patients. Furthermore, during infection, patients with lung cancer are more prone to die than patients with other cancer types. The study also showed higher possibilities of the cancer survivors to have severe conditions than normal patients[58]. The outcomes of a multi-cancer study support the previous findings. However, the pro-inflammatory conditions due to the cytokine storm can raise the possibility of malignancies apart from the diseases like ARDS. Pro-inflammatory mediators, such as IL6, TLRs, and TNF-α, help in tumorigenesis. Specifically, IL6 works to protect the cancer cell from DNA damage, apoptosis, etc. Attenuation of IL6 is considered as therapeutic technique for malignancies. From multiple studies, we have observed that the upregulation of the IL6 can be one of the roads towards the malignancies from SARS-CoV-2 infection. In fact, SMAD3 protein, which is known as a prime tumor suppressor, has been attenuated during cytokine storm[59]. This also enhances the possibilities of multiple malignancies. More elaborately, SMADs, and especially SMAD3, are responsible for the TGF-mediated immune suppression, which would later promote the favorable conditions for metastasis. Therefore, systematic expression of SMAD3 proteins can maintain the activity of the TGF-β signaling. So far it has been observed that the cross-talk between TGF-β signaling and different pathways is responsible for different cancers. For example, the cross talk between IRS-1 signaling pathway and TGF-β signaling in colon cancer or a cross talk between the BCL signaling pathway and TGF-β signaling in hepatocellular carcinoma. Therefore, the attenuated levels of SMAD3 may enhance the chances of the malignancy development, especially at the post-COVID situation. Importantly, two proteins (TMPRSS2 and FURIN), which are exclusively associated with the viral uptake, are directly and partially androgen-modulated. Mollica et al.[60] showed that the promoter of the TMPRSS2 can be modulated via androgen receptors (AR). Lucas et al. showed how the androgen-mediated protease TMPRSS2, which plays an important role in viral uptake, might help in creating the proteolytic cascade during prostate carcinoma and prostatic neoplasms. Therefore, the androgen levels can be the determinant factor to understand the viral vulnerability, which is two times in male. However, the amount of the androgen hormone varies among the genders. More specifically, males and postmenopausal females, who have elevated androgen levels, are more vulnerable and susceptible to the infection. Among the disease classes, different types of malignancies are associated with maximal number of host cell proteins affected by SARS-CoV-2 (shown in Figure 6). This number is alarming not only during the infection but also in the post-COVID situation.

SARS-CoV-2 and its Impact on Neurodegenerative and Neuropsychiatric Diseases

In the case of neurodegenerative and neuropsychiatric diseases, the disease possibilities have been increased due to the neuroinflammation mediated by SARS-CoV-2 infection. As per previous evidence, neuroinflammatory responses are very common and usual for disease, such as Alzheimer's disease (AD), Parkinson's disease (PD), multiple sclerosis (MS), etc. due to the intervention of the pro-inflammatory mediators, such as chemokines, cytokines, interleukins, etc. Cytokine storm at CNS can affect the microglial cells, a specialized population of macrophages. Usually, these cells show a stable and inactive immunophenotype in healthy brain. However, functional dysregulation of the microglial cells can occur due to the unusual interactions with cytokines and also due to exposure to soluble Amyloid-beta ($A\beta$) (during pathogenic progression of dementia in AD). On the other hand, microglial cells are associated with the phagocytosis of extracellular $A\beta$. These could be the reasons behind increasing the rate of AD. Though, the relation between immunology and neurodegenerative and neuropsychiatric diseases are not linear. The similar type of neuroinflammation is also observed during the cellular deposition of the α-synuclein in PD. Basically, the cytokine storm during the intrusion of SARS-CoV-2 may disrupt blood-brain barrier and may be associated with early progression of the aforementioned diseases[11]. Also, few cases of schizophrenia have also been noticed in post-COVID population[61]. Though the reasons for this increase in schizophrenia incidences are not clear. However, the relation between viral infection and neuropsychiatric diseases is well known. Some of the previous researches have shown that chronic viral infection is responsible for loosing cognitive senses[62,63]. SARS-CoV-2 can follow the same path, as Rogers et al. have summarized in their recent review[64]. As per the review report, multiple case studies have shown that neuropsychiatric dysfunctions, including anxiety, insomnia, depression, impaired memory, etc. are associated with 7%-41% of the COVID-19 cases, whereas the overall percentage of affected patients is around 63%. Two possibilities can be discussed in this regard. First, the genetic disorders, where the infection can work as a dominator[58]. Second, neuroinflammation can be another reason behind it.

Methods

In this study, the Multiple Sequence Alignment (MSA) is performed for each selected family of from Pfam database[65]. Subsequently, we analyzed the sequences in evolutionary time-frame, and designed individual structures for each class of SARS-CoV-2 proteins by applying in silico techniques. The detailed flowchart of the utilized framework is shown in Figure 2.

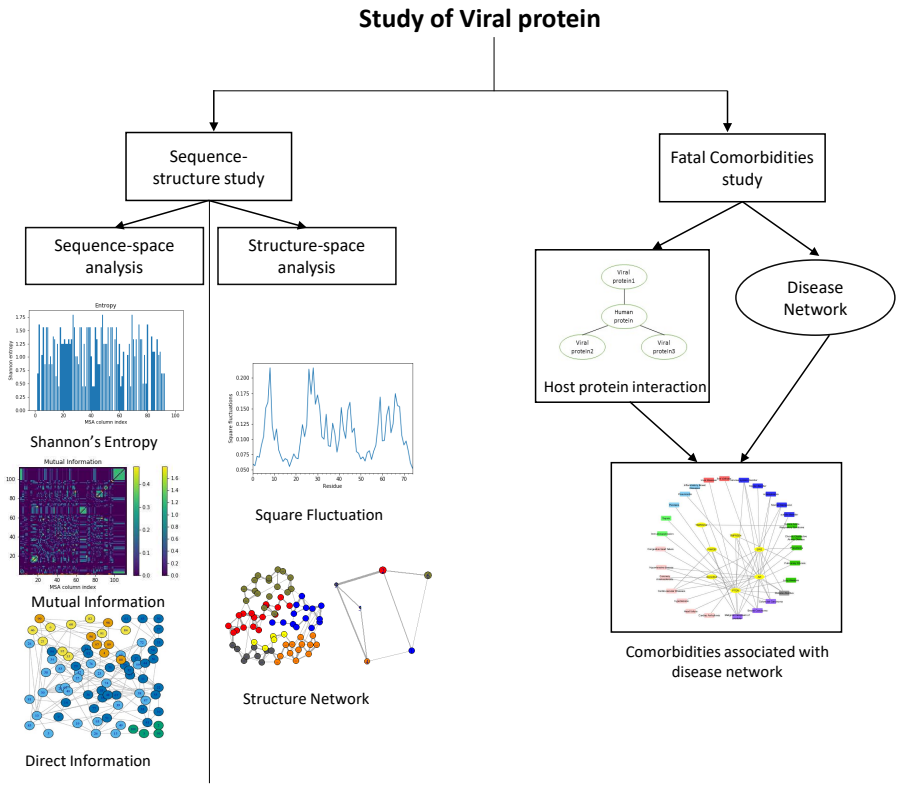

Figure 9. The flowchart of the proposed method

Data Description and Corresponding Sources

The data for this experiment are curated from multiple data repositories. We have mostly focused on protein sequences, structures and host cells associated diseases. Data have been fetched accordingly. All the data sources are listed in Table 1.

Serial Number	Database	Usage
1	Pfam	The sequences of the selected protein families are curated to perform Multiple Sequence Alignment (MSA)
2	UniProt	15 host cell proteins are considered which are responsible for CoVID-19 reported in this database
3	RCSB PDB	The information of each proteins is fetched from this database to execute the structure network analysis
4	DisGeNET	A disease list is prepared to establish the disease network
5	Chakrabarty et. al	A disease list is prepared to establish the disease network

Table 1. The data repositories used to fetch the data and utilized in this study are listed in this table.

Hydropathy Calculation

Hydrophobicity of amino acids is a measure that reflects their solubility in water. In order to identify the hydrophobic regions in a protein, the well-known Kyte-Doolittle scale[66] was used. During this calculation, a protein sequence is scrutinized with a sliding predefined window. At each position, the mean hydropathy value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window. As it is advised that the peptide length should be greater than double the window size to get any useful information from the hydropathy plot, and as we are interested in the identification of the surface-exposed hydrophilic sites, the window size was set to 7.

Sequence-Structure Space Study

Here we have two stages of sequence space analysis and one stage of the structure space analysis to study the variability of the viral proteins. The sequence space analysis has been sub divided in two parts, such as Shannon's Entropy analysis and Direct Coupling Analysis. In structure space, we used the PDB structures to construct corresponding structure networks. The detail method is given below.

Data Pre-Processing for Sequence Space

For the sequence-based analysis, we have analyzed 17 distinct families for five individuals (shown in Supplementary Table T1). A hidden Markov model (HMM)-based MSA was performed, where the aligned sequences were stored as $MS^n_{aligned}$, here n represents the protein class. Two exclusive entropy method were applied to $MS^n_{aligned}$ sequentially. The detailed method is described below.

Residual Occupancy and Shannon's Entropy

To understand the frequency of amino acids at different positions within the amino acid sequence, we have applied Shannon's Entropy (SE) scoring[33]. From the MSA, the frequency of amino acid substitutions at specific position within the amino acid sequence was studied. Low and high SE scores represent conservation or randomness of amino acids at a certain position, respectively.

$$S(j) = -\sum_{b_j=1}^{20} P(b_i) log(b_i) \tag{1}$$

In equation 1, j represents the sequence position, the probability of amino acid b to be present at the jth column of MSA have been represented by $P(b_j)$. $S(j)$ represents the SE score.

Coupling Analysis

Sequence conservation provides information in the single-dimensional space. However, the conservation of physico-chemical properties of each protein can be studied based on the evolutionary co-varying residues at certain positions within the amino acid sequence. In this regard, we applied two distinct scoring systems, such as Mutual Information (MI) and Direct Information (DI). The input for both scoring equations is MSA of the individual protein family. MI shows rate of co-variation of two patches in the presence of other residues, whereas DI is special case of MI, where co-varying patches are disentangled from the rest of the sequence. Therefore, co-evolution is nothing but evolutionary correlation between two positions within the sequence. High correlation represents stronger connection.

Calculation of MI is shown in Equation 2, where $P_{ij}(A,B)$ is considered as a joint probability, and $P_i(A)$ and $P_j(B)$ are individual probabilities of the residues at the positions i and j. The resultant MI scores represent the coupling strength among two amino acid positions/residues.

$$MI_{ij} = \sum_{AB} P_{ij}(A,B) log \frac{P_{ij}(A,B)}{P_i(A)P_j(B)} \tag{2}$$

$$DI_{ij} = \sum_{AB} P_{ij}^{(dir)}(A,B) ln \frac{P_{ij}^{(dir)}(A,B)}{P_i(A)P_j(B)} \tag{3}$$

In equation 3, DI has been described. where $P_{ij}^{(}dir)$ represents reweighted frequency where the coupled pairs are disentangled from the rest of the MSA set.

Graph Theoretical Modelling and Eigenvector Community Detection

The depiction of the physico-chemical properties from the coupling analysis can be shown through a weighted graph of G_{DI}, where $(V_{DI}, E_{DI}) \in G_{DI}|V_{DI}$ represented residues and E_{DI} denoted weighted edges between directly correlated coupled pairs. In this article, our aim was to study the evolutionary dynamics of the proteins associated with SARS-CoV-2. In this regard, eigenvector-based community detection has been applied on G_{DI}. Usually, eigenvector centrality of a node in a network analyzes the strength of its connectivity with the remaining nodes. Following the concept, community based on eigenvectors represents the dynamic connectivity among the nodes. Therefore, community detection in G_{DI} can provide modules with higher rate of co-varying residues, where the residues are densely connected.

Root Mean Square Fluctuation

Root Mean Square Fluctuation (RMSF) is measure of the particle deviation. In RMSF, a mean over time is considered for a residue j at the current position and some reference position. The definition of the RMSF is given in Equation 4.

$$RSMF_j = \left(\frac{1}{T} \sum_{t_k=1}^{T} mod(r_j(t_k) - r_j^{re})^2 \right)^{.05} \tag{4}$$

Where T is time over which the mean has been taken for reference position of the particle j , r_j^{re}. The RMSF has been observed based on the reference position of the particle j over time.

Normal mode-based Structure Network Analysis

Proteins are dynamic in nature and their fluctuations play vital roles in their functions. To understand the sequential orchestration, a protein structure network was established. The network is constructed based on the Normal mode analysis (NMA), which works better for large structural rearrangements. In this network diagram, amino acids are represented as nodes and their strength of non-covalent interactions are depicted through edges. Equation 5 is used to establish the interaction.

$$F_{pq} = \left[\frac{X_{pq}}{\sqrt{X_p * X_q}} \right] * 100 \geqslant F_t \tag{5}$$

Here, X_{pq} is the number of side chain, p and q are atom pairs of residues. X_p and X_q are the normalization factor for residue type p and q[67,68]. F_t is the threshold of interaction strength whereas 4% is the default value.

In this study, a cross-correlation matrix was calculated depending on the correlation matrix of NMA. The hypothesis behind the NMA of protein is the vibrational normal modes manifesting the lowest frequencies, which unfold the largest protein movements and the functionally relevant ones. The tertiary structures for the particular proteins of different families are generated from I-TASSER[69]. NMA provides a comprehensive outlook of protein tertiary structure without coarse-gaining.

Firstly, NMA controls Cartesian coordinates as independent variables. NMA can also represent the chain connectivity of polypeptide chains. Subsequently, activity of the parameters can easily be controlled tweaking dihedral bonds. Along with that, NMA considers the individual movement of each of the residues, which perhaps helps defining the external and internal chemistry comprehensively. The value of cross-correlation is represented by the weight connections of a particular node. Simultaneously, a full residue network is established based on the correlation network analysis. Girvan-Newman clustering method is used with threshold of 0.3 to split it into densely correlated coarse-grained community cluster network[70].

Disease Network

The 15 host cell proteins were considered, which are responsible for the COVID-19 reported in UniProt (www.uniprot.org). To identify the patient's vulnerability towards COVID-19 and the diseases, which may originate due to the current pandemic disease, a disease network was established for each selected protein. In this regard, a human-based PPI network was built using STRING database to capture the effects of viral invasion on fatal comorbidities. The interacting neighbors of the acquired protein were selected based on the homology, experimentally validated interactions, and co-expression. Further, a disease list was prepared by considering DisGeNET[71] and data reported by Chakrabarty et. al[72]. The selected diseases are divided based on the mostly affected organs due to infection. The categorization of diseases are cardiovascular diseases, skin related problems, disease of mental health, immunity, viral diseases, respiratory problems, kidney diseases, Gastrointestinal system, Glucose metabolism disease, carcinomas, etc[73]. Finally, the resultant diseases from each of the samples were compared with the prepared disease list. The common diseases between two lists were considered to be more influential and were shortlisted accordingly.

Competing interests

The Authors have no competing interests.

References

1. Goh, G. K. M., Dunker, A. K. & Uversky, V. N. Understanding Viral Transmission Behavior via Protein Intrinsic Disorder Prediction: Coronaviruses. *J. Pathog.* **11**, 1–13 (2012).
2. Uversky, V. N. A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Sci.* **22**, 693–724 (2013).
3. Giri, R. *et al.* Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses. *Cell. Mol. Life Sci.* 1–34 (2020).
4. Bianchi, M. *et al.* Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics? *BioMed Res. Int.* 1–6 (2020).
5. Puelles, V. G. & et al. Multiorgan and Renal Tropism of SARS-CoV-2. *The New Engl. J. Medicine* **383**, 590–592 (2020).
6. Calabrese, F. & et al. Pulmonary pathology and COVID-19: lessons from autopsy. The experience of European Pulmonary Pathologists . *Virchows Arch.* **477**, 359–372 (2020).
7. Naicker, S., Yang, C. W., S. J. Hwang and, B. C. L., Chen, J. H. & Jha, V. The Novel Coronavirus 2019 epidemic and kidneys. *Kidney Int.* **97**, 824–828 (2020).
8. Abenavoli, L., Gentile, I., Maraolo, A. E. & Negro, F. SARS-CoV-2 and liver damage: a possible pathogenetic link. *Hepatobiliary Surg. Nutr.* **9**, 322–324 (2020).
9. Bandyopadhyay, D. & et al. COVID-19 Pandemic: Cardiovascular Complications and Future Implications. *Am. J. Cardiovasc. Drugs* **20**, 311–324 (2020).
10. Elrashdy, F., Redwan, E. M. & Uversky, V. N. Intrinsic disorder perspective of an interplay between the renin-angiotensin-aldosterone system and SARS-CoV-2. *Infect. Genet. Evol.* **85** (2020).
11. Song, P., Li, W., Xie, J., Hou, Y. & You, C. Intrinsic disorder perspective Cytokine storm induced by SARS-CoV-2. *Clin. Chimica Acta* **509**, 280–287 (2020).
12. Roche, L. D. & Mesta, F. Oxidative Stress as Key Player in Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) Infection. *Arch. Med. Res.* **51**, 384–387 (2020).
13. Olsen, J. G. & K. Teilum and, B. B. K. Behaviour of intrinsically disordered proteins in protein–protein complexes with an emphasis on fuzziness. *Cell. Mol. Life Sci.* **74**, 3175–3183 (2017).

14. Wu, C. & et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sinica B* **10**, 766–788 (2020).

15. Xue, B. *et al.* Structural Disorder in Viral Proteins. *Chem. Rev.* **114**, 6880–6911 (2014).

16. Goh, G. K. M., Dunker, A. K., Foster, J. A. & Uversky, V. N. Rigidity of the Outer Shell Predicted by a Protein Intrinsic Disorder Model Sheds Light on the COVID-19 (Wuhan-2019-nCoV) Infectivity. *Biomolecules* **10**, 331 (2020).

17. Dey, A., Sen, S., Uversky, V. N. & Maulik, U. Structural facets of POU2F1 in light of the functional annotations and sequence-structure patterns. *J. Biomol. Struct. Dyn.* 1–13 (2020).

18. Sen, S., Dey, A., Chowdhury, S., Maulik, U. & Chattopadhyay, K. Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis. *BMC bioinformatics* **19**, 231–242 (2020).

19. Mitra, R. *et al.* Decoding critical long non-coding rna in ovarian cancer epithelial-to-mesenchymal transition. *Nat. communications* **8**, 1–12 (2017).

20. Liu, Y. & Bahar, I. Sequence Evolution Correlates with Structural Dynamics. *Mol. Biol. Evol.* **29**, 2253–2263 (2012).

21. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. United States Am.* **108**, E1293–E1301 (2011).

22. Chowdhury, S. *et al.* Evolutionary Analyses of Sequence and Structure Space Unravel the Structural Facets of SOD1. *Biomolecules* **9**, 1–18 (2019).

23. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinforma.* **40**, 1–8 (2008).

24. Yang, J. *et al.* The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

25. Pinero, J. & et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants . *Nucleic Acids Res.* **45**, D833–D839 (2017).

26. Uversky, V. N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front. Phys.* **7**, 1–18 (2019).

27. Kutyshenko, V. P. & et al. Solution structure and dynamics of the chimeric SH3 domains, SHH- and SHA-"Bergeracs". *Biochimica et Biophys. Acta* **1794**, 1813–1822 (2009).

28. Li, L., Uversky, V. N., Dunker, A. K. & Meroueh, S. O. A computational investigation of allostery in the catabolite activator protein. *J. Am. Chem. Soc.* **129**, Journal of the American Chemical Society (2007).

29. Xue, B., Li, L., Meroueh, S. O., Uversky, V. N. & Dunker, A. K. Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol. BioSystems* **5**, 1688–1702 (2009).

30. Melnik, T. N., Povarnitsyna†, T. V., Glukhov, A. S., Uversky, V. N. & Melnik, B. S. Sequential Melting of Two Hydrophobic Clusters within the Green Fluorescent Protein GFP-cycle3. *Biochemistry* **50**, 7735–7744 (2011).

31. Melnik, B. S. *et al.* SS-Stabilizing Proteins Rationally: Intrinsic Disorder-Based Design of Stabilizing Disulphide Bridges in GFP. *J. Biomol. Struct. Dyn.* **29**, 815–824 (2012).

32. Melnika, B. S., Molochkovb, N. V., Prokhorova, D. A., N.Uverskycd, V. & P.Kutyshenko, V. Molecular mechanisms of the anomalous thermal aggregation of green fluorescent protein. *Biochim Biophys Acta* **1814**, 1930–1939 (2011).

33. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).

34. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P. & Dunker, A. K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61**, 176–182 (2005).

35. Peng, K. *et al.* Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinforma. Comput. Biol.* **3**, 35–60 (2005).

36. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).

37. Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding . *Nucleic Acids Res.* **46**, W329–W337 (2018).

38. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophys. Acta* **1804**, 996–1010 (2010).

39. Schoeman, D. & Fielding, B. C. Coronavirus envelope protein: current knowledge. *Virol. J.* **16**, 1–22 (2019).

40. Dyson, H. J. Roles of Intrinsic Disorder in Protein-Nucleic Acid Interactions. *Mol. Omics* **8**, 97–104 (2012).

41. Wang, C., Uversky, V. N. & Kurgan, L. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* **16**, 1486–1498 (2016).

42. Wu, Z. *et al.* In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. *FEBS Lett.* **589**, 2561–2569 (2010).

43. Mishra, P. M., Verma, N. C., Rao, C., Uversky, V. N. & Nand, C. K. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. *Prog. Mol. Biol. Transl. Sci.* **174**, 1–78 (2020).

44. Goh, G. K. M., Dunker, A. K., Foster, J. A. & Uverskye, V. N. Shell disorder analysis predicts greater resilience of the SARS-CoV-2 (COVID-19) outside the body and in body fluids. *Microb. Pathog.* **174**, 1–6 (2020).

45. Tanaka, T., Kamitani, W., DeDiego, M. L., Enjuanes, L. & Matsuura, Y. Severe Acute Respiratory Syndrome Coronavirus nsp1 Facilitates Efficient Propagation in Cells through a Specific Translational Shutoff of Host mRNA. *J. Virol.* **86**, 11128–11137 (2012).

46. Thuaud, F., Ribeiro, N., Nebigil, C. G. & Desaubry, L. Prohibitin Ligands in Cell Death and Survival: Mode of Action and Therapeutic Potential. *Chem. & Biol.* **20**, 316–331 (2013).

47. Zhang, H., Penninger, J. M., Li, Y., Zhong, N. & Slutsky, A. S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine volume* **46**, 586–590 (2020).

48. Zhou, P., Yang, X. L. & Shi, Z. L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

49. Vandelli, A. *et al.* Structural analysis of SARS-CoV-2 andpredictions of the human interactome. *Prog. Mol. Biol. Transl. Sci.* 1–30 (2020).

50. Davies, J. P., Almasy, K. M., McDonald, E. F. & Plate, L. Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies. *bioRxiv* 1–44 (2020).

51. Lucas, J. M. & et al. The Androgen-Regulated Protease TMPRSS2 Activates a Proteolytic Cascade Involving Components of the Tumor Microenvironment and Promotes Prostate Cancer Metastasis. *Cancer Discov.* **4**, 1310–1325 (2014).

52. Ming, Y. & Qiang, L. Involvement of Spike Protein, Furin, and ACE2 in SARS-CoV-2-Related Cardiovascular Complications. *SN Compr. Clin. Medicine* **114**, 1–6 (2020).

53. Li, S. W. *et al.* SARS coronavirus papain-like protease induces Egr-1-dependent up-regulation of TGF-β1 via ROS/p38 MAPK/STAT3 pathway. *Sci. Reports* (2016).

54. Wang, C. Y. & et al. SARS coronavirus papain-like protease up-regulates the collagen expression through non-Samd TGF-β1 signaling. *Virus Res.* **235**, 58–66 (2017).

55. Zhao, X., Nicholls, J. M. & Chen, Y. G. Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-beta signaling. *The J. Biol. Chem.* **283**, 3272–3280 (2007).

56. Montopoli, M. & et al. Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532). *Annals Oncol.* **31**, 1040–1045 (2020).

57. Magro, G. SARS-CoV-2 and COVID-19: Is interleukin-6 (IL-6) the 'culprit lesion' of ARDS onset? What is there besides Tocilizumab? SGP130Fc. *Cytokine: X* **2**, 100029 (2020).

58. Liang, W. & et al. Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *The Lancet Oncol.* **21**, 335–337 (2020).

59. Gurram, R. K., Kujur, W., Maurya, S. K. & Agrewala, J. N. Caerulomycin A Enhances Transforming Growth Factor-β (TGF-β)-Smad3 Protein Signaling by Suppressing Interferon-γ (IFN-γ)-Signal Transducer and Activator of Transcription 1 (STAT1) Protein Signaling to Expand Regulatory T Cells $(Tregs)^*$. *J. Biol. Chem.* **289**, 17515–17528 (2014).

60. Mollica, V., Rizzo, A. & Massari, F. The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer. *Futur. Oncol.* **21**, 1–5 (2020).

61. Singal, C. M. S., Jaiswal, P. & Seth, P. SARS-CoV-2, More than a Respiratory Virus: Its Potential Role in Neuropathogenesis. *ACS Chem. Neurosci.* **11**, 1887–1899 (2020).

62. Huang, J. *et al.* Potential of SARS-CoV-2 to Cause CNS Infection: Biologic Fundamental and Clinical Experience. *Front. Neurol.* **11**, 1–9 (2020).

63. Sun, X. & et al. Cytokine storm intervention in the early stages of COVID-19 pneumonia. *Cytokine Growth Factor Rev.* **53**, 38–42 (2020).

64. Rogers, J. P. *et al.* Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic. *Lancet Psychiatry* **7**, 611–627 (2020).

65. Finn, R. D. & et al. Pfam: the protein families database) Infectivity. *Nucleic Acids Res.* **42**, D222–D230 (2014).

66. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

67. Brinda, K. V. & Vishveshwara, S. A Network Representation of Protein Structures: Implications for Protein Stability. *Biophys. J.* **89**, 4159–4170 (2005).

68. Kannan, N. & Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method . *J. Mol. Biol.* **292**, 441–464 (1999).

69. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).

70. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. United States Am.* **99**, 7821–7826 (2002).

71. Pinero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).

72. Chakrabarty, B., Das, D., Bulusu, G. & Roy, A. Network-Based Analysis of Fatal Comorbidities of COVID-19 and Potential Therapeutics. *ChemRxiv* 1–27 (2020).

73. Wu, F., Zhao, S., Yu, B. & et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

Figures

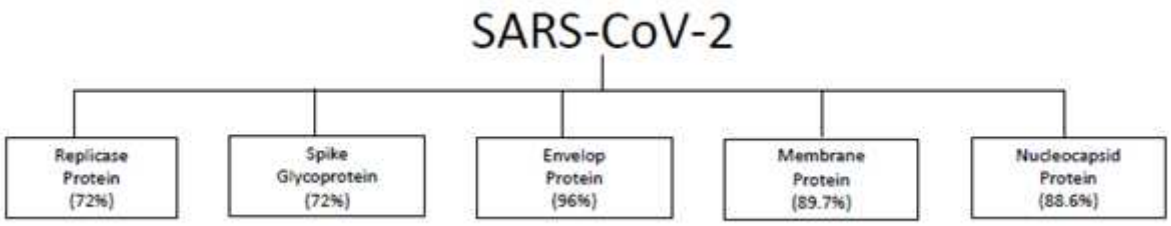

Figure 1

Sequence similarity of the selected proteins of SARS-CoV-2 and SARS-CoV

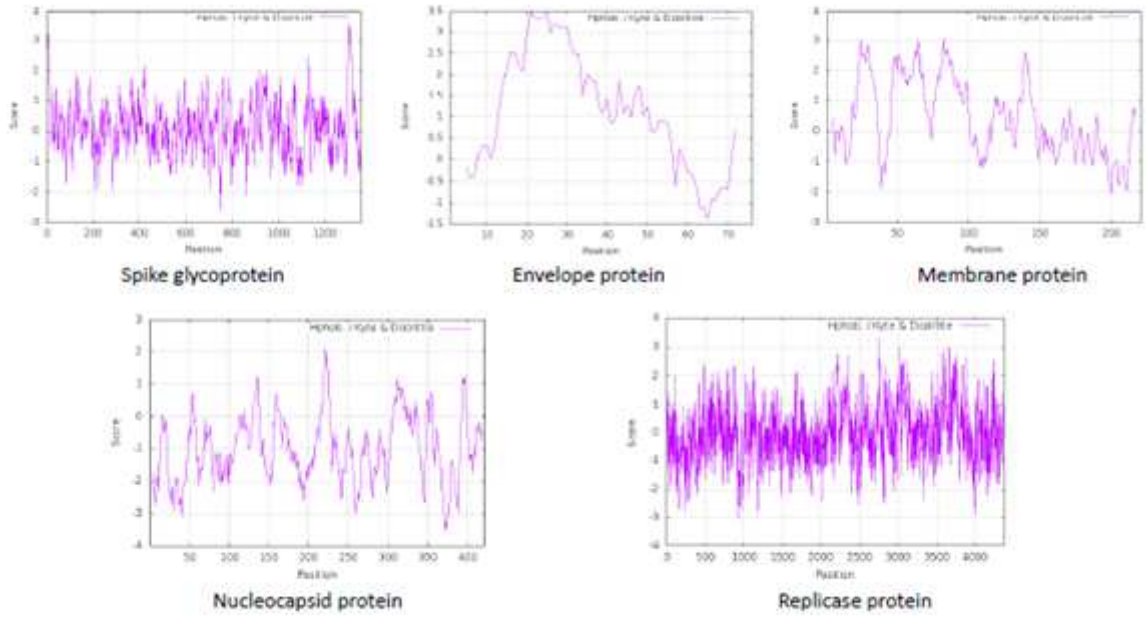

Figure 2

The two-dimensional hydropathy plot produced for the selected proteins based on the Kyte and Doolittle scale of amino acid hydropathy.

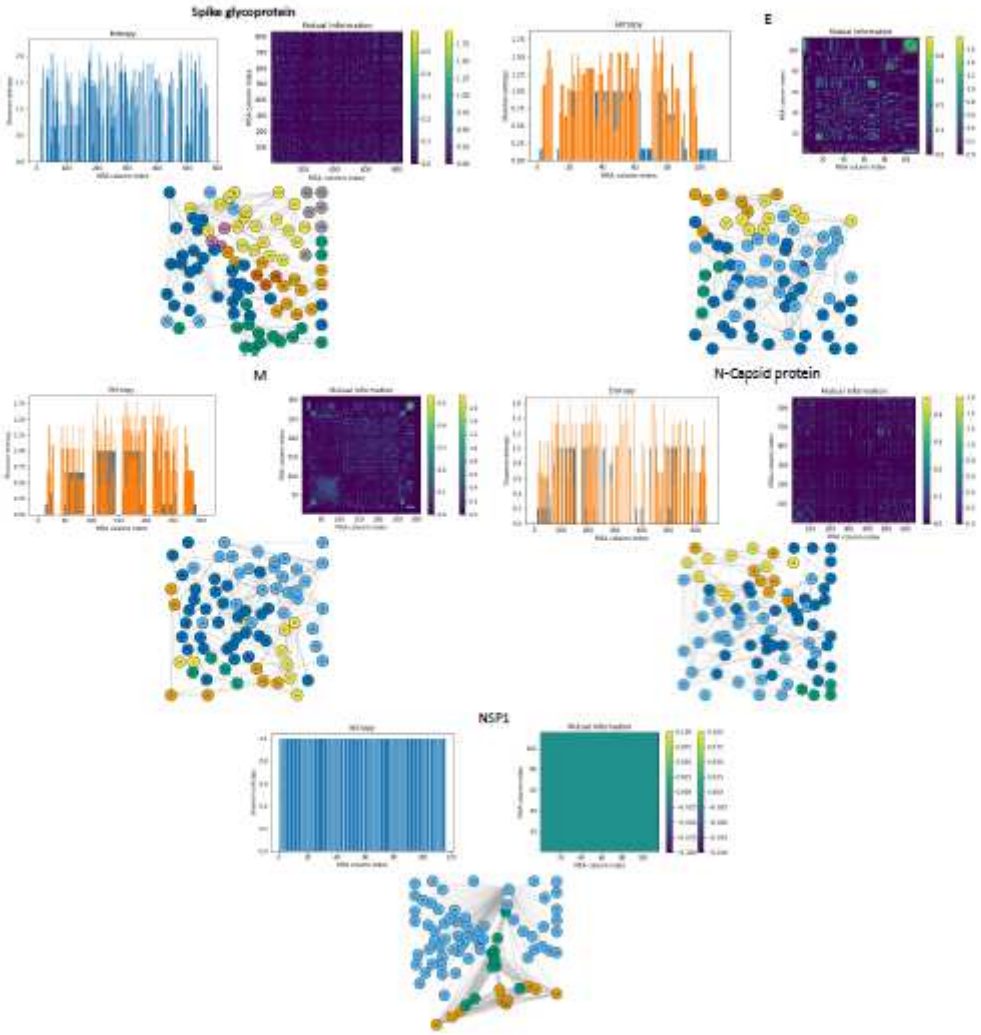

Figure 3

The sequence space analysis is performed for the selected protein families along with the Shannon Entropy calculation, coupling analysis, and community detection techniques

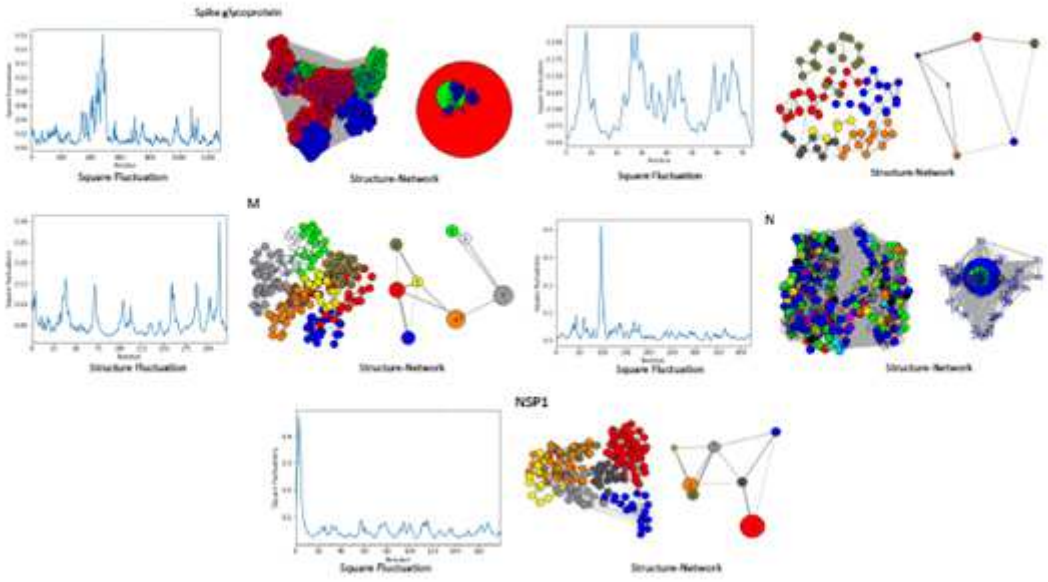

Figure 4

In the structure space analysis of the selected protein families the structure fluctuations and structure network are performed.

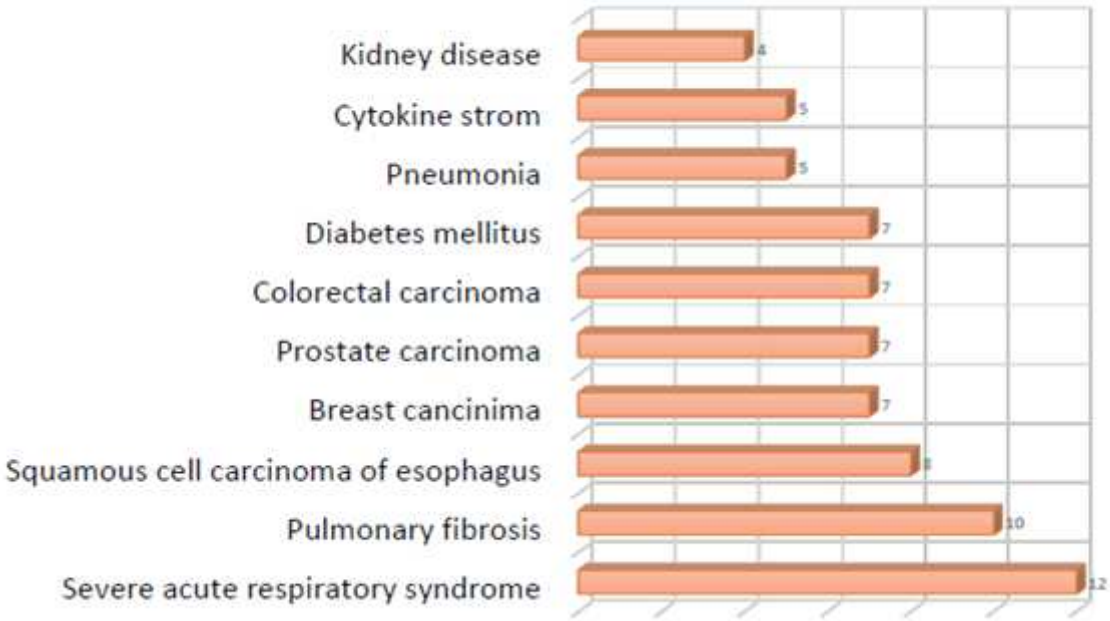

Figure 5

To understand the involvement of the proteins with the distortion of the organ behavior we performed a Venn diagram of the disease shared by the proteins.

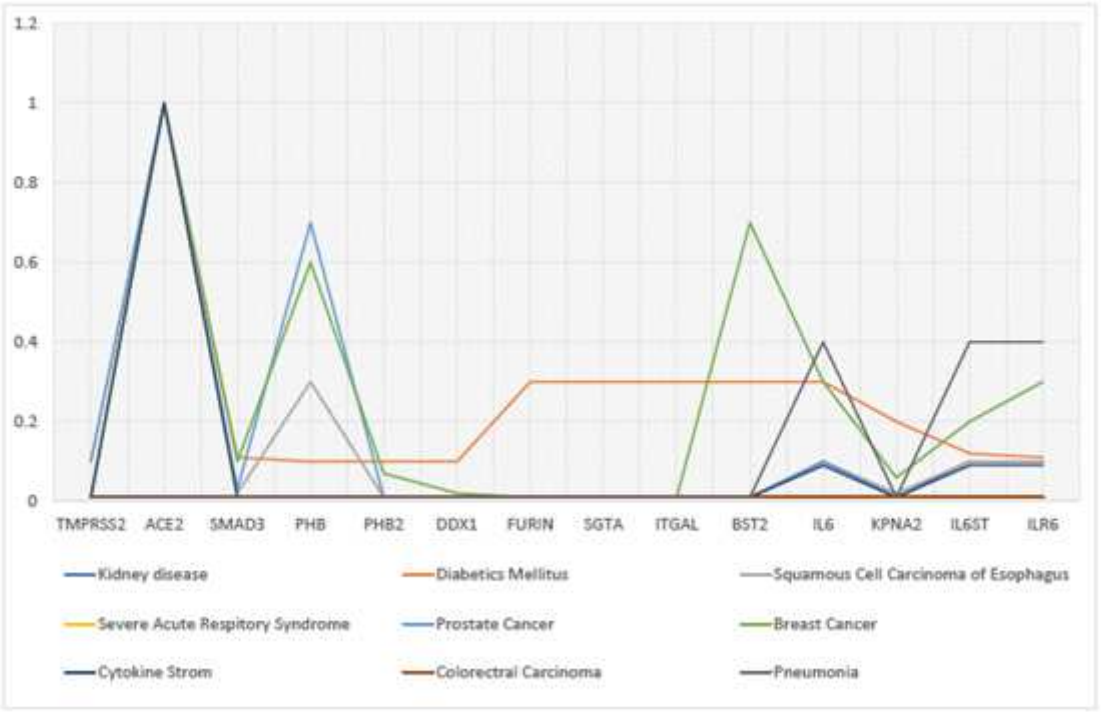

Figure 6

The dependency score of the diseases with their respective protein are considered and a line graph is constructed. Each line of the graph represents a particular disease marked in the figure.

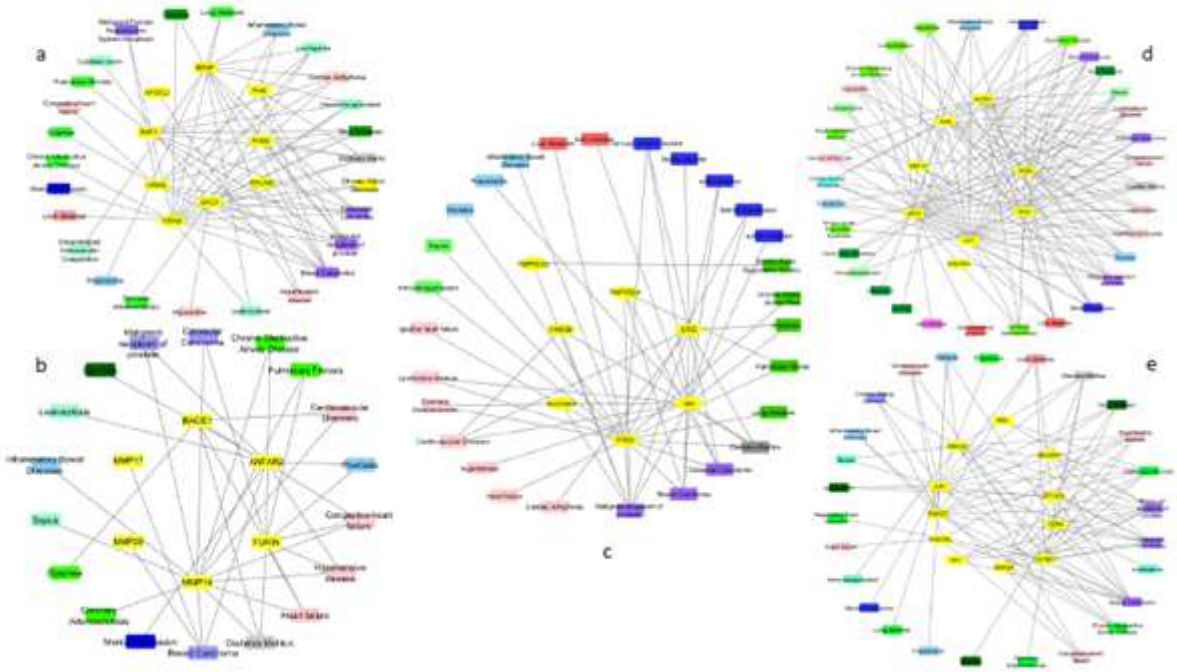

Figure 7

Five proteins a. PHB , b. FURIN , c. TMPRSS2, d. ACE2, e. SMAD3 are shown based on a common disease among the 14 proteins.

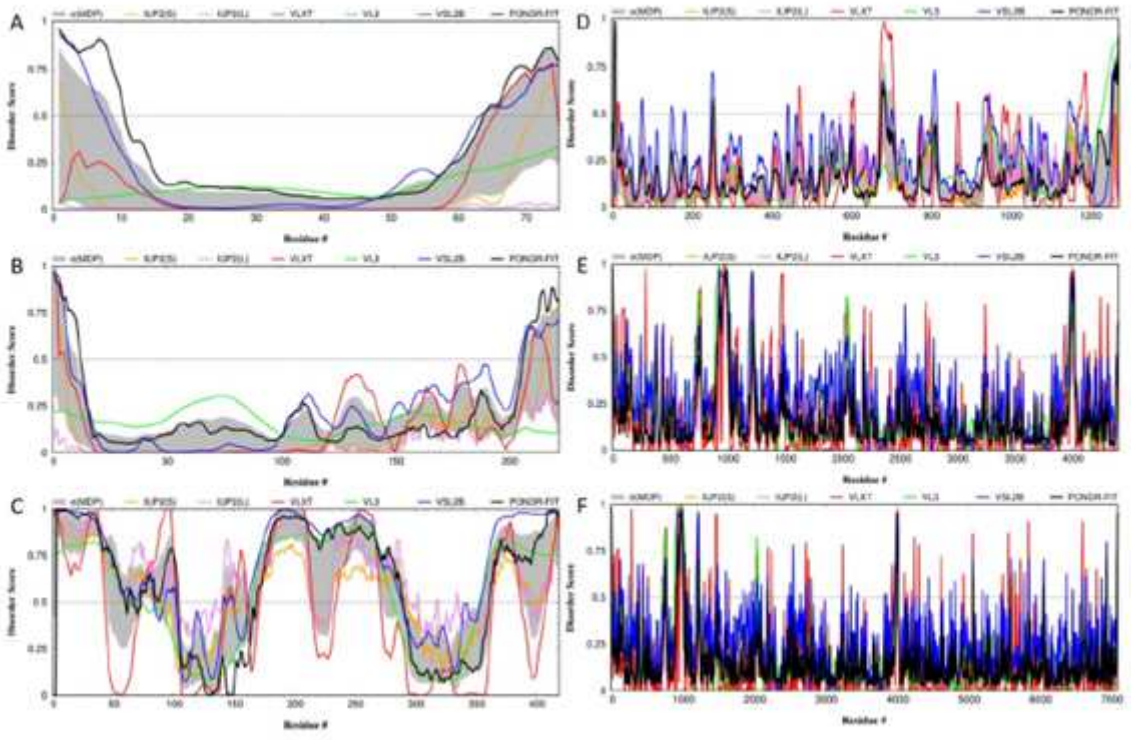

Figure 8

Evaluation of intrinsic disorder predisposition of major SARS-CoV proteins: envelope (A), membrane (B), nucleocapsid (C), spike (D), ORF1a (E) and ORF1ab (F). These profiles were generated using DiSpi web crawler that aggregate the results from a number of well-known disorder predictors: PONDR R VLXT33, PONDR R VSL234, PONDR R VL335, IUPred short and IUPred long36, 37, and PONDR R FIT38

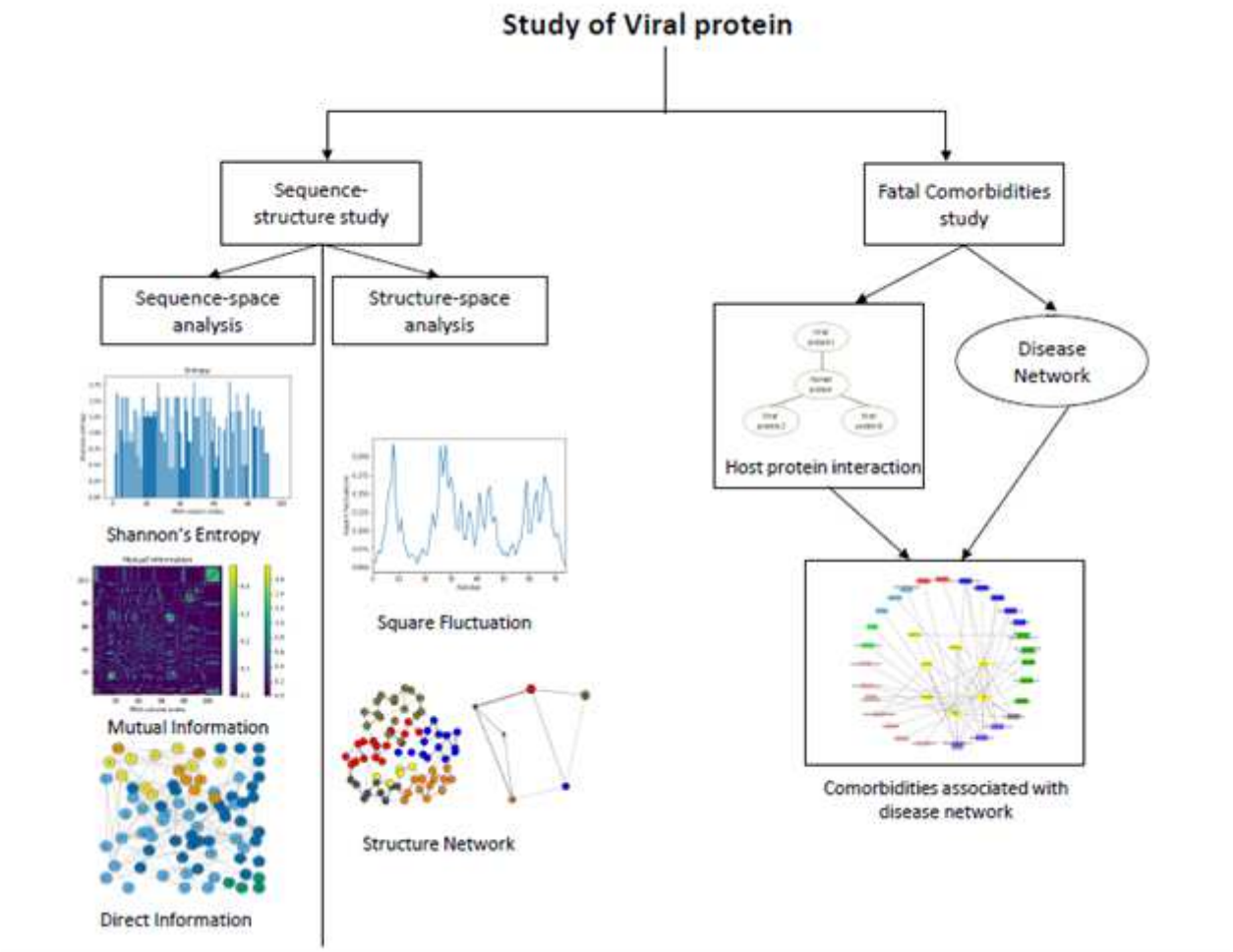

Figure 9

The flowchart of the proposed method

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1.pdf
- S2.pdf
- S3.pdf
- T1.xlsx